OUTLOOK
Brain cancer

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



STELLAR OUTBURST

3D simulations reveal source of variability in massive blue stars

PAGE 498

UK SCIENCE

DEAL OR NO DEAL

Perspectives on Brexit six months before the split

PAGE 452

MEDICAL RESEARCH

WHAT'S IN A NAME?

Time to halt the scourge of mesenchymal stem cells

PAGE 455

BIOCHEMISTRY

ROLL OUT THE β-BARREL

Targeted protein binding by computational design

PAGES 471 & 485

NATURE.COM
27 September 2018
Vol. 561, No. 7724

THIS WEEK

EDITORIALS

HISTORY The lasting scholarly value of a physical archive **p.434**

WORLD VIEW Scrap rankings and treat all authors as equal **p.435**



MISCONDUCT Food researcher Brian Wansink resigns after investigation **p.439**

The costs of climate inaction

A new analysis breaks down the likely social cost of carbon emissions by country, and should make unhappy reading for politicians.

Politics, according to the nineteenth-century German statesman Otto von Bismarck, is the art of the next best. The global approach of politicians to tackling climate change is a sorry example of this.

The problem: destructive storms that hit the United States and south-east Asia this month are the latest reminder of how vulnerable societies across the world are to climate extremes. The best political solution might seem to be to subordinate all policies — domestic and international — to the goal of stabilizing Earth's climate. This is difficult. So, instead, the world must rely on the effectiveness of voluntary actions that nations have agreed on under a non-binding international compromise treaty forged in Paris in 2015.

For all its symbolic power, that Paris treaty is a truly second-best solution. Even if it had worked as advertised, the promised cuts in greenhouse-gas emissions are weak. And now the withdrawal of the United States — and, de facto, of Australia — has substantially weakened the global consensus before the treaty has even come into effect.

Discussions on how and when it will start will resume at a two-week United Nations meeting in December in Katowice, Poland. Those attending would do well to read a study published this week in *Nature Climate Change* that highlights just how irrational it is for the politicians who represent many large economies to settle for next best (K. Ricke *et al. Nature Clim. Change* http://doi.org/ct7x; 2018).

The analysis revisits the concept of the social cost of carbon: the cumulative economic impact of global warming caused by (or attributed to) each tonne of the pollutant sent into the atmosphere. This study goes a step further than previous ones and estimates the likely cost to different countries. In doing so, it reveals the countries projected to take the hardest hits.

China and the United States, the world's two largest emitters of carbon dioxide, will incur some of the highest social costs of carbon of all countries, the scientists report, with respective estimated impacts of US$24 per tonne and $48 per tonne. India, Saudi Arabia and Brazil also feature towards the top. In these countries — unlike in Canada, northern Europe and Russia — temperatures are already above the economic optimum. And climate-induced damage increases with wealth and economic growth, meaning that more-valuable property might sit in harm's way.

Combined country-level costs (and benefits) add up to a global median of more than $400 in social costs per tonne of CO_2 — more than twice previous estimates. On the basis of CO_2 emissions in 2017, that's a global impact of more than $16 trillion. The new analysis is based on a set of climate simulations, rather than a single climate model, and the authors calculated future harm using empirical damage functions that were independently developed for that purpose.

The revised costs are still ballpark figures, based on relatively uncertain assumptions on climate physics, emission trajectories, socio-economic development and climate-driven economic damage. In fact, climate change could also have impacts on international trade, security and human migration that calculations of the social costs of carbon don't capture. But the concept is valuable, nonetheless. Acting like a magnifying glass, it highlights horrendous climate-impact inequality. For example, whereas Canada and Russia are still gaining economic benefits worth up to $10 per tonne of CO_2 from rising temperatures, India is already paying an exorbitant price ($86 per tonne).

It also shows that the way in which society currently prices carbon (as a means of reducing its use and protecting future generations) is an order of magnitude too low. The current price of carbon on the European market is just over $20. And in most other parts of the world, it's effectively zero.

"For all its symbolic power, the Paris climate treaty is a truly second-best solution."

The new analysis sends a powerful message from a future that most people say they want to avoid. In response, will politicians up their ambition and aim for the best — and necessary — solution? The paper unfortunately comes too late to be included in the special report from the Intergovermental Panel on Climate Change on the effects of 1.5 °C in global warming, due to be published next month. But it adds to the growing body of research that unpicks that global effect, and breaks it down into regions and countries. This will be needed to plan mitigation and also to prepare for adaptation.

One government that should pay particular attention to the latest work is that of the United States, where the social cost of carbon has been taken into account in policymaking — for example, in car standards. President Donald Trump's advisers have previously challenged cost estimates used by the US Environmental Protection Agency as being too high. The revised calculations suggest that the opposite is the case. ■

Deal making

European science is already suffering from the damaging effects of Brexit.

So this is how the United Kingdom's relationship with the European Union ends: not with a bang, or even a whimper, but with a series of technical notices published quietly on the website of Her Majesty's Government.

The series of briefings — the latest batch was released earlier this month — discuss the possible consequences should Britain fail to agree terms with the EU on how to remove itself from the bloc. In those circumstances — the 'no deal' scenario — Britain would be ejected from a raft of shared laws and regulations, including those governing the free movement of people, goods and services across borders in the EU. With regulatory systems on either side of the English Channel out of step, experts have warned, the worst-case scenario could see

© 2018 Springer Nature Limited. All rights reserved.

chaos and disruption to supply chains, transport and daily life.

Scientists are among those who have been anxiously scanning the government notices. The documents include predictions of the effects on research funding (bad), access to satellite-navigation systems (minimal) and warnings about dangerous space debris (cross fingers and hope for the best). Government spokespersons have been at pains to play down the negatives highlighted by their own analyses, but in each case the attempt at reassurance has been the same: 'It won't come to that. We're trying very hard to agree a deal.'

Officials need to do so in just six months: the two-year period since the United Kingdom gave its formal notice to quit the EU expires on 29 March 2019. Most politically pressing is to find a way to distinguish between Ireland (which will remain in the EU) and Northern Ireland (which won't) without erecting a hard border, which, at worst, could reignite violence. But question marks hang over a string of issues, including how the United Kingdom should manage its nuclear research outside the EU, and whether the import of scientific equipment and reagents will be affected.

A sensible assessment of the situation says that the consequences of no deal are simply so bad that neither Britain nor the EU will let it happen. A compromise will surely emerge: either an extension of the deadline or some kind of holding commitment to make agreements in the near future. But numerous obstacles remain, among them that the ruling Conservatives will have to secure a vote in Parliament, and many of the party's hard-liners are in no mood to compromise.

Some sectors are rightly making arrangements for a no-deal scenario. The UK Office for Nuclear Regulation, for example, says it is training staff and developing the IT infrastructure needed to work outside Euratom, the EU umbrella body. And some UK universities are strengthening links with overseas institutions in the hope that this will keep them plugged into European funding streams.

Regardless of whether or not a deal is done, many scientists are already seeing and feeling the impact of Brexit, as we report in a News Feature this week (page 452). Although it might seem on the surface that it is business as usual until key decisions are made, science and scientists in Britain are suffering as a result of the uncertainty. Researchers are less likely to get collaborators on projects, because academics in Europe view them as a risky bet and are teaming up with universities elsewhere. Some are finding it harder to fill key positions. Others feel unable to apply for EU funding, and the country is losing its reputation as an international hub of excellent research. Many scientists are feeling tired and disappointed. The uncertainty is taking a personal and emotional toll.

"Although it might seem that it is business as usual, science and scientists are suffering."

Some UK scientists do see opportunities. Earlier this month, plant scientists pointed out that a no-deal Brexit could spare them from new and controversial moves in Brussels to classify gene-editing techniques as genetic modification, and so subject to all the same strict rules. That might be good for them, but it also reinforces a broader concern about the future of EU policy. On issues from regulation of genetically modified crops to allowing research with embryonic stem cells, the UK government has historically been more bullish than other European nations, and this has helped to forge the continent into a world-leader in many fields. Without Britain's contribution as a moderating and rational voice on key decisions, Europe's attitude to science will suffer.

On this point, the EU can take some concrete steps to keep Britain at the table. UK officials will no longer be able to serve on advisory panels after Brexit, but some 100 UK scientists also work in Brussels in second-tier positions, such as for the Joint Research Centre, which informs EU legislation and regulations in policy areas from environment to migration. As things stand, they will be expected to leave with Brexit. Allowing them to stay on would be a small but pragmatic way for the EU to ease the impact of Britain's departure. More must be sought. A united Europe is a major force in global research. It will be less of one after the United Kingdom goes. ■

In the archives

The discovery of Galileo's long-lost letter highlights the value of physical repositories.

Modern scholars don't always have to physically visit museums and archives around the world to seek secrets of the past. Many collections have been digitized, and much can be done with these online resources. But can anything beat the thrill of being there and finding an item assumed lost to history? That's what happened last month at the London archives of the Royal Society, with the discovery of a letter of great historical importance (see page 441).

Written by Galileo Galilei in 1613, the letter sets down for the first time the scientist's gripes with the Vatican's doctrine on astronomy. His forthright objections launched one of science history's most famous battles, which culminated in the Inquisition's condemnation of Galileo for heresy 20 years later. Different copies of the letter had circulated, and their content has been tirelessly analysed and discussed by historians. But seeing the original for the first time, with its scorings-out and word substitutions, solves a long-standing mystery about whether a version sent to the Inquisition in Rome had been doctored — and, if so, by whom.

Galileo, it now seems clear, doctored his original letter himself, to make the language less aggressive, as soon as he realized the trouble heading his way. This suggests that the editing was not the malign work of theologians trying to make a stronger case against him, as had been assumed by the nineteenth-century scholar Antonio Favaro, whose 20-volume *The Works of Galileo Galilei* is a main reference work.

Discovering an old document that allows a gap in history to be filled is a rare event in the life of a science historian. It makes all those years spent in dusty archives — or squinting at digital archives on a screen — worthwhile. The 1613 Galileo letter could have been found by anyone, given that it was hiding in plain sight in the Royal Society's online catalogue. So the discovery happened by chance, made by a visiting Italian scholar who was filling the last hour of his working day with an unplanned browse. Spotting a reference in the online archive, with mounting excitement, he asked to see it.

Perhaps no scientist in history has been as deeply studied as has Galileo, a prodigious scribe who is widely considered the father of the scientific method. There has been enough analysis of surviving copies of Galileo's letters, documents and books for some scholars not to have been overly surprised that the great scientist might indeed have rewritten a little of his own history. Scholars who have pored over his works for decades, and who understand the context of his life, his personality and turns of phrase, have a feel for these things. But seeing the editing in Galileo's own handwriting adds certainty to the interpretation. And just having the object is itself a tangible cultural gain.

There are many ways to piece history together. Research into the lives of famous people such as Galileo drives much of the knowledge we have about the past. By contrast, the Venice Time Machine (see *Nature* **546**, 341–344; 2017), a massive project to digitize a 1,000-year archive and apply machine-learning techniques, promises to dig out knowledge about the lives of the non-famous. Offline or online, scholarly analysis or machine learning, all of these approaches combine to build a more complete perspective.

Digital resources are of inestimable value to historians, but the discovery of the Galileo letter underlines the need to protect original objects, many of them stored in vulnerable museums and libraries. So does the devastating loss of artefacts in the fire at Brazil's National Museum in Rio de Janeiro earlier this month. We will never know if an equivalent to Galileo's letter perished in the flames there. Some history has been lost. But some, if we can preserve it, is merely waiting to be discovered. ■

© 2018 Springer Nature Limited. All rights reserved.



COMMERCIALIMAGE.NET

No more first authors, no more last authors

If we really want transdisciplinary research, we must ditch the ordered listing of authors that stalls collaborative science, says **Gretchen L. Kiser**.

Every academic scientist has heard a tale of someone being shafted on an authorship list, or had it happen to them. Less appreciated is how much the attribution of credit impedes cross-disciplinary approaches to difficult questions. It creates a negative feedback loop that hinders research.

Most scientists agree that research questions and approaches have become more complex, so the need to engage in expanded team science has increased. I've found, however, that there is great reluctance among faculty members to join such efforts. I find myself asking, 'What if we completely blow up the way in which we attribute authorship?' I suspect that if we got rid of first authors, last authors and the fight for asterisks, we might interrupt the negative feedback loop and see more innovation.

Since 2012, I've led the Research Development Office at the University of California, San Francisco (UCSF). One of our goals is to bring together researchers of varying backgrounds to encourage innovative thinking and new approaches. My team identifies and cajoles 'champions' to invite colleagues to participate in team-building events. We offer financial and logistical support; we bring in interesting speakers; we provide drinks and food (and not just pizza!) — all to get scientists to talk to each other about their research, needs and ambitions. But the resource that really matters is not mine to dispense: credit for scientific contributions.

WHAT IF WE COMPLETELY BLOW UP THE WAY IN WHICH WE ATTRIBUTE AUTHORSHIP?

There are real successes: one of our 'speed-networking' events at UCSF introduced neurologist Dena Dubal, who investigates the molecular mechanisms of longevity and neurodegenerative disease, to psychologist Aric Prather, who researches the effects of stress on health. That led to a project that revealed an association between chronic psychological stress and lower levels of a longevity hormone. They published that work and continue to collaborate (A. A. Prather *et al. Transl. Psychiatr.* **5**, e585; 2015).

Other teams we've helped have received follow-on support from external funders such as the US National Institutes of Health. Surveys tell me that faculty members enjoy our team-building events, even when they did not expect to, and that they would recommend them to others.

Nevertheless, there seems to be an undeclared disincentive for researchers to build unconventional collaborations. I get frustrated with the disconnect between what we say about the need for transdisciplinary teams to solve complex problems and the reluctance to try something new to build those teams.

The assessment of publications during promotion and tenure decisions is a big part of the problem. Although these processes often have some mechanism to recognize a researcher's team contributions, the culture remains largely unchanged from 50 years ago. The gravitas associated with 'first' and 'senior' authorship is entrenched. What about the middle author who might have significantly altered the approach? Or the fourth-place author who linked different disciplines? Often these researchers are left to find only self-satisfaction.

Many journals now allow, and even require, statements that explain contributors' roles in their publications. Taxonomies and standardized vocabularies for describing authors' roles have been developed. Similarly, promotion and tenure committees are using contribution narratives in their assessments. These changes are helping. They capture a fuller spectrum of a researcher's productivity so that evaluators can consider more than where someone sits in an author list.

Still, I've had senior faculty members tell me that, even though they look at the contribution narratives, they still expect to see first-author and then senior-author papers when assessing candidates.

Meanwhile, research projects are starting to incorporate data that no one on the immediate team collected, and there are no settled conventions for crediting outside researchers or incentivizing that valuable work.

We need a cultural shift to recognize and reward scientists who make their work useful to others, including researchers who might never meet but whose data are used. One way to make this happen is to get rid of ordered author lists. By developing author contribution taxonomies and narratives, we have already acknowledged the need to reflect the multifaceted nature of authorship. Large consortia and organizations are adopting contribution frameworks to reflect author roles and participation more accurately. We are also moving to use repository tools that assign authorship to different types of research output, such as data sets. More effort, creativity and diversity of thought are needed. We should stop trying to apply old attribution models to the innovative ways we now generate data.

If we can reveal the shape of proteins at atomic resolutions, tweak genes to order and detect cosmic signals from the beginning of time, then surely we can work out better ways to represent author contributions. We already send complex basic research and clinical data into 'information commons' and build computational 'knowledge network' tools to inform patient diagnostics and therapeutics. A well-annotated data set might be combined with other data to expand its impact synergistically. Can we imagine an author attribution method that would use cutting-edge computational tools similar to those being applied to scientific research itself? A tool that gives credit where credit is due?

If we acknowledge the products of research in more-innovative ways, the value of 'team-ness' might grow in academic culture and the cutting edge will get sharper. Perhaps, then, I won't have to cajole anyone to participate in team-building activities. ■

Gretchen L. Kiser *is executive director of the Research Development Office at the University of California, San Francisco.*
e-mail: gretchen.kiser@ucsf.edu

© 2018 Springer Nature Limited. All rights reserved.

SEVEN DAYS *The news in brief*

SPACE

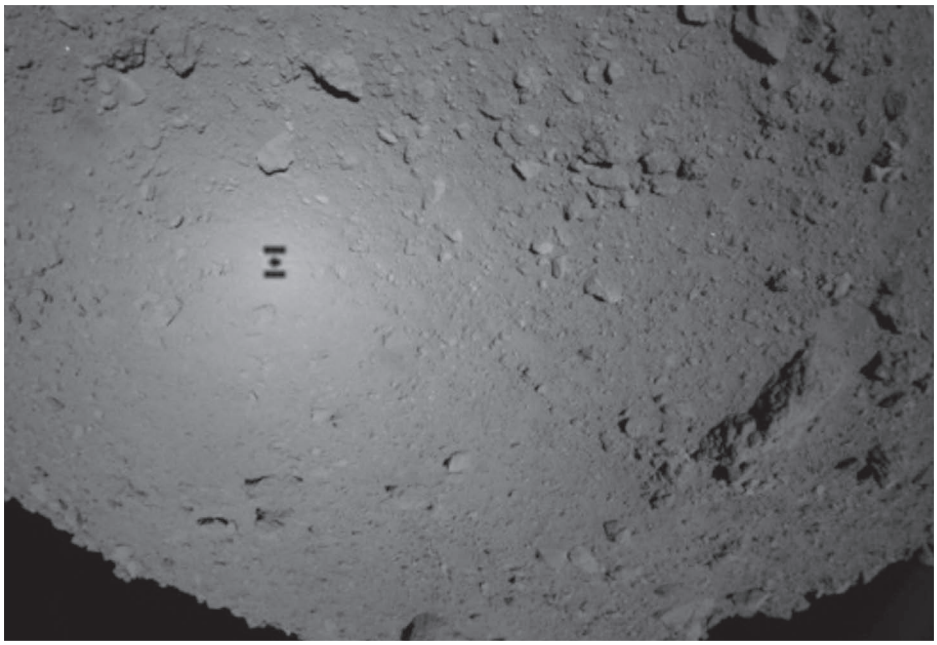
Gravity probe glitch

A US–German gravity-mapping mission that launched in May has not gathered science data since mid-July because of a problem with a key instrument. The twin satellites of the Gravity Recovery and Climate Experiment Follow-On (GRACE-FO) mission measure minute differences in the distance between them to chart changes in Earth's gravity. Doing so can reveal the distribution and flow of water and ice across the planet. But part of a microwave instrument aboard one of the satellites shut itself down on 19 July following signs of an electrical problem. Mission managers plan to switch to a back-up system later this month, said Frank Webb, the mission's project scientist at NASA's Jet Propulsion Laboratory in Pasadena, California, on 19 September.

POLICY

Clinical-trial fines

The US Food and Drug Administration (FDA) is proposing heavy fines for pharmaceutical companies that fail to report clinical-trial results online. In a 20 September draft guidance statement, the FDA said that failure to register trials or submit results to the government database ClinicalTrials.gov could result in penalties of up to US$10,000 per day. The agency would send a notification letter to the company warning of an infraction. The company would then have 30 days to comply before facing fines. The public has 60 days to submit comments on the proposed rule. Studies have found that about 58% of industry-funded trials and 61% of trials funded by the US National Institutes of Health in the United States did not report their results five years after completion.



JAXA

Japan mission makes asteroid history

Japan's asteroid mission Hayabusa2 has become the first to land moving rovers on the surface of an asteroid. On 22 September, the Japan Aerospace Exploration Agency (JAXA) confirmed that the mission's twin rovers, called MINERVA-II 1A and 1B, had landed safely on the space rock Ryugu, and that both were moving on the surface. The Hayabusa2 mothership temporarily dropped close to the surface (pictured) of the 1-kilometre-wide asteroid to deploy the probes last week. Mission controllers lost communication with the rovers in the hours after deployment, when the landers were on the asteroid's far side, as seen from the orbiter. But the hexagonal rovers, which hop around slowly on the low-gravity surface, have now sent back their first blurry colour images. In October, the mothership is scheduled to touch down to collect a sample, which it will return to Earth. Scientists hope that studying the asteroid will help them to understand the origins of the Solar System's planets.

US biodefence

The US government has revised its plans for responding to biological threats. On 18 September, the White House released the first US biodefence strategy that spans multiple government agencies. In another first, the plan covers not only deliberate bioterror threats, but also naturally occurring outbreaks and infectious diseases that escape a lab accidentally. The White House has also created a steering committee to coordinate operations in the event of a biological attack or a major disease outbreak; it will be led by the Department of Health and Human Services. It will undertake an immediate review of agencies' biodefence strategies and capabilities in order to identify any gaps.

RESEARCH

Agriculture data

A coalition including the Food and Agriculture Organization of the United Nations (FAO), the Bill and Melinda Gates Foundation and national governments has launched a US$500-million effort to help developing countries gather data on farmers and promote rural development. The initiative, announced on 24 September, will scale up surveys developed separately by the FAO and the World Bank that gather information on agriculture efforts around the world. The programme aims to fill a yawning information gap about what hundreds of millions of farmers living

© 2018 Springer Nature Limited. All rights reserved.

in poverty are doing across 50 countries in Africa, Asia and Latin America.

PEOPLE

Scientist cleared

An Indian space scientist, Nambi Narayanan, has been granted compensation after being falsely accused of trading secrets with a foreign country. On 14 September, India's Supreme Court condemned the original police investigation against Narayanan and awarded 5 million rupees (US$68,600) to the former Indian Space Research Organisation scientist. In 1994, he was arrested on charges that he sold secret information on cryogenic rocket engines to Pakistan. Although India's Central Bureau of Investigation and a local court cleared Narayanan of the allegations in 1996, the case was reopened by local authorities, which led to a protracted legal battle between various government agencies. The Supreme Court has now established a committee to investigate the police officers who arrested Narayanan.

Misconduct case

Food psychologist Brian Wansink (**pictured**) has submitted his resignation to



LEFTY SHIVAMBU/GALLO IMAGES/GETTY

Cornell University in Ithaca, New York, and will retire at the end of the academic year, according to a 20 September statement from the university. The move comes as Cornell announced the findings of an internal investigation that concluded that Wansink had committed academic misconduct. The university's year-long investigation found that the scientist had misreported research data, used "problematic" statistical techniques and failed to document his results properly. Cornell's statement added that Wansink had been removed from teaching and research responsibilities, and would spend the remainder of the academic year cooperating with its ongoing review of his work. Wansink, who is known for his research on how external factors such as packaging influence what people eat, said that "there was no fraud, no intentional misreporting, no plagiarism, or no misappropriation" in an e-mailed statement to *Nature*.

FUNDING

Funds for children

The US Department of Health and Human Services (HHS) plans to take money from research programmes to help house immigrant children whose parents were detained after entering the United States illegally. The HHS, which includes the US National Institutes of Health (NIH) and the Centers for Disease Control and Prevention (CDC), sent a letter to members of a Senate spending panel that oversees the agency, saying that it plans to transfer US$87.3 million from the NIH and $16.7 million from the CDC for the housing programme. The costs of housing these children have ballooned since April, when President Donald Trump instituted a policy that has led to the separation of many families, as part of a crackdown on illegal immigration. The letter was first reported on 19 September by Yahoo News.

HEALTH

Tuberculosis trends

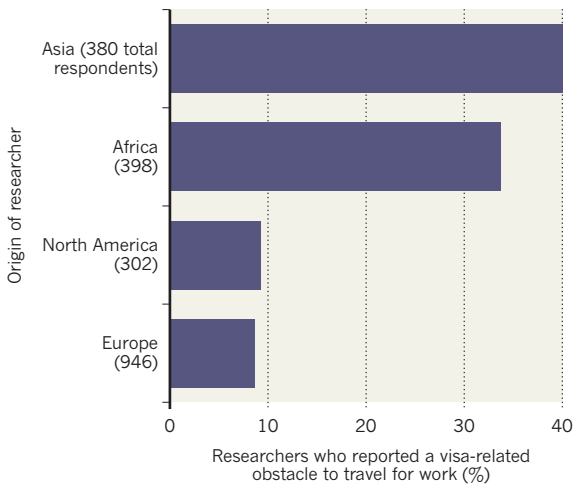
The incidence of tuberculosis (TB) is falling globally by about 2% per year, the World Health Organization said in a report on 18 September. Between 1.5 million and 1.7 million people are estimated to have died from TB in 2017, and 9 million to 11 million people developed the disease. India has the largest TB burden of any country, accounting for more than one-quarter of the world's cases. Nearly one-tenth of new cases globally occurred in people with HIV, who are particularly susceptible to infection because HIV weakens their immune systems. Multi-drug-resistant tuberculosis is rising globally, with 161,000 cases in 2017 compared with 153,000 in 2016.

TREND WATCH

Researchers from Africa and Asia encounter much greater obstacles in getting visas to visit other countries for work than do researchers from Western nations. A global survey of academics' movements found that Asian researchers were more than four times, and African academics more than three times, as likely as European or North American researchers to report visa-related obstacles to travelling for work. The survey polled 2,465 researchers in 109 countries, including 494 in Africa and 457 in Asia, and was conducted by the RAND Corporation for the Wellcome Trust, a research-funding charity in London, and the Together Science Can Campaign, which promotes international collaboration. The most common complaints globally involved the length of time needed to process a visa request, the complexity of application forms, costs associated with making an application and a lack of clarity on rules and procedures. Around 45% of researchers with an African nationality reported travelling to another country for work only very rarely, compared with around 17% of European researchers.

VISA TROUBLES

Researchers from Africa and Asia are most likely to experience visa problems when travelling for work, particularly for short visits.



SOURCE: RAND CORPORATION

© 2018 Springer Nature Limited. All rights reserved.

NEWS IN FOCUS

FUNDING Palaeontologist loses £1-million grant after bullying investigation **p.442**

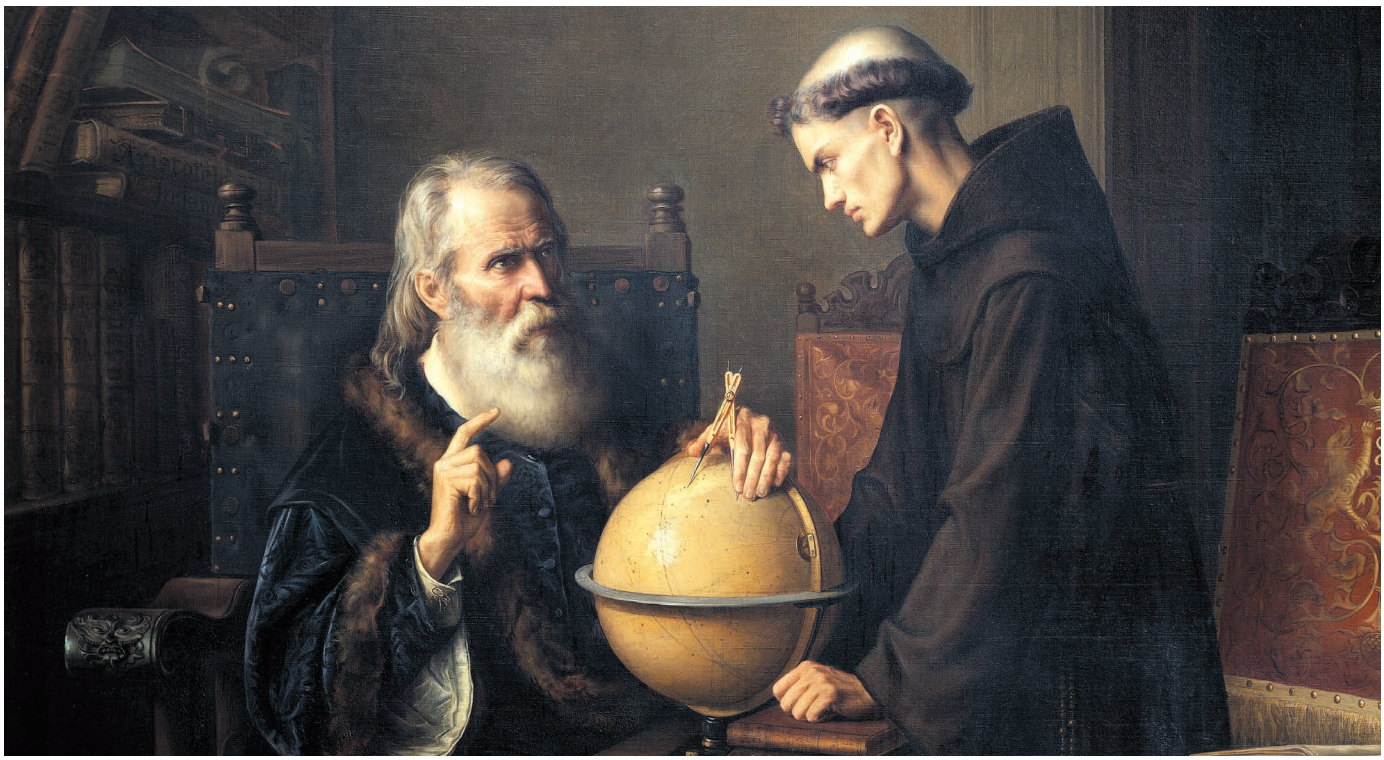
ECOLOGY Deep-sea surveys reveal vibrant life in ocean zone marked for mining **p.443**

ARCHAEOLOGY Prehistoric kids laboured as salt miners and brickmakers **p.445**



BREXIT How researchers are preparing for the approaching split **p.452**

DEAGOSTINI/GETTY/PAINTING BY FELIX PARRA



Galileo's celestial ideas prompted a trial by the Inquisition in Rome, which convicted him on "vehement suspicion of heresy" in 1633.

SCIENCE HISTORY

Lost Galileo letter reveals he tried to dodge Inquisition

Chance find shows astronomer softened his heretical claims — then lied about his edits.

BY ALISON ABBOTT

It had been hiding in plain sight. The original letter — long thought lost — in which Galileo Galilei first set down his arguments against the church's doctrine that the Sun orbits Earth has been discovered in a misdated library catalogue in London. Its unearthing and analysis expose critical new details about the saga that led to the astronomer's condemnation for heresy in 1633.

The seven-page letter, written to a friend on 21 December 1613 and signed "G.G.", provides the strongest evidence yet that, at the start of his battle with the religious authorities, Galileo actively engaged in damage control and tried to spread a toned-down version of his claims.

Many copies of the letter were made, and two differing versions exist — one that was sent to the Inquisition in Rome and another with less inflammatory language. But because the original letter was assumed to be lost, it wasn't clear whether incensed clergymen had doctored the letter to strengthen their case for heresy — something Galileo complained about to friends — or whether Galileo wrote the strong version, then decided to soften his own words.

Galileo did the editing, it seems. The newly unearthed letter is dotted with amendments — and handwriting analysis suggests that Galileo wrote it. He shared a copy of this softened version with a friend, claiming it was his original, and urged him to send it to the Vatican.

The letter has been in the Royal Society's possession for at least 250 years, but escaped the notice of historians. It was rediscovered in the library there by Salvatore Ricciardo, a science historian at the University of Bergamo in Italy, who visited on 2 August for a different purpose, and browsed the online catalogue.

"I thought, 'I can't believe that I have discovered the letter that virtually all Galileo ▶

© 2018 Springer Nature Limited. All rights reserved.

scholars thought to be hopelessly lost,'" says Ricciardo. "It seemed even more incredible because the letter was not in an obscure library, but in the Royal Society library."

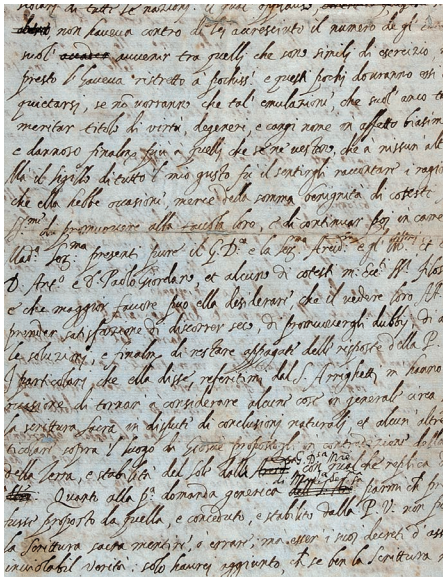
Ricciardo and his colleagues Franco Giudice at the University of Bergamo and Michele Camerota of the University of Cagliari in Italy describe the letter's details and implications in an article in press at the Royal Society journal *Notes and Records*. Some science historians declined to comment on the finding before they had scrutinized the article. But Allan Chapman, a science historian at the University of Oxford, UK, and president of the Society for the History of Astronomy, says: "It's so valuable — it will allow new insights into this critical period."

MIXED MESSAGES

Galileo wrote the 1613 letter to Benedetto Castelli, a mathematician at the University of Pisa in Italy. In it, Galileo set out for the first time his arguments that scientific research should be free from theological doctrine.

He argued that the scant references in the Bible to astronomical events should not be taken literally, because scribes had simplified these descriptions so that they could be understood by common people. Religious authorities who argued otherwise, he wrote, didn't have the competence to judge. Most crucially, he reasoned that the heliocentric model of Earth orbiting the Sun, proposed by Polish astronomer Nicolaus Copernicus 70 years earlier, is not actually incompatible with the Bible.

Galileo, who by then was living in Florence, wrote thousands of letters, many of which are scientific treatises. Copies of the most significant were made by different readers and widely circulated. His letter to Castelli caused a storm.



THE ROYAL SOCIETY

The letter shows amendments in Galileo's hand.

Of the two versions known to survive, one is now held by the Vatican Secret Archives. This version was sent to the Inquisition in Rome on 7 February 1615, by a Dominican friar named Niccolò Lorini. Historians know that Castelli then returned the original 1613 letter to Galileo, and that on 16 February 1615 Galileo wrote to his friend Piero Dini, a cleric in Rome, suggesting that the version Lorini had sent to the Inquisition might have been doctored. Galileo enclosed with that letter a less inflammatory version, which he said was the correct one, and asked Dini to pass it on to Vatican theologians.

His letter to Dini complains of the "wickedness and ignorance" of his enemies, and lays out his concern that the Inquisition "may be in part deceived by this fraud which is going around under the cloak of zeal and charity". At least a dozen copies of the version Galileo sent to Dini are now held in different collections.

Beneath its scratchings-out, the signed copy discovered by Ricciardo shows Galileo's original wording — and it is the same as in the Lorini copy. The changes are telling. In one case, Galileo referred to certain propositions in the Bible as "false if one goes by the literal meaning of the words". He crossed through the word "false", and replaced it with "look different from the truth". In another section, he changed his reference to the Scriptures "concealing" its most basic dogmas, to the weaker "veiling".

This suggests that Galileo moderated his own text, says Giudice. To be certain that the letter really was written in Galileo's hand, the three researchers compared individual words in it with similar words in other works written by Galileo around the same time.

CHANCE DISCOVERY

The historians are now trying to trace how long the letter has been at the Royal Society, and how it arrived there. In the catalogue, the document was dated 21 October 1613, but when Ricciardo examined it, he saw it was actually dated 21 December 1613. That might be one reason why the letter has been overlooked by Galileo scholars, says Giudice. The letter was included in an 1840 Royal Society catalogue — but was also misdated there, as 21 December 1618.

For now, the researchers are stunned by their find. "Galileo's letter to Castelli is one of the first secular manifestos about the freedom of science — it's the first time in my life I have been involved in such a thrilling discovery," says Giudice. ■

FUNDING

Scientist stripped of grant

Dinosaur discoverer disciplined by University of Bath, UK, after investigation into bullying.

BY HOLLY ELSE

A research-funding foundation has revoked a £1-million (US$1.3-million) grant from prominent palaeontologist Nicholas Longrich, who was disciplined by his institution, the University of Bath, UK, after an investigation found that he had breached its anti-harassment policy.

Longrich was part of a team that in 2015 reported the first four-legged fossil snake, a high-profile discovery published in *Science*[1]; in 2010, he grabbed the media spotlight with his discovery of the whimsically named *Mojoceratops perifania* dinosaur[2].

The Leverhulme Trust awarded the £998,185 grant — at least three-quarters of which was dedicated to paying research assistants and postgraduate students — to Longrich in 2016 for research on a mass-extinction event that marked the end of the Cretaceous era 66 million years ago.

"We can confirm that Dr Longrich's grant has been withdrawn but his doctoral students will not be disadvantaged by this," said a spokesperson for the foundation, which distributes £80 million of research funding each year. Leverhulme declined to add more details, and referred further queries to the University of Bath.

On 23 August, the university told *Nature* in a statement that it had received a formal complaint about Longrich in late May 2018 relating to a "potential breach" of the university's dignity and respect policy, which aims to prevent bullying, harassment and victimization of students and staff.

In early June, it began a formal investigation, which ended in July. "The investigation panel considered written and oral statements, taking evidence from the complainant, the subject of the complaint and a number of others," said the university. "The conclusion reached was that though there had been no malicious intent, the formal complaint should be upheld."

The university issued Longrich with an oral warning and made changes to his "supervisory arrangements" with current students. It said

© 2018 Springer Nature Limited. All rights reserved.

that the changes will apply to future students, too.

The university declined to give more details of the findings of the investigation. In its statement, it said: "We are providing further information only where we are satisfied that the privacy of individual students and staff would not be compromised and the necessary consents have been obtained."

On 19 September, after Leverhulme's decision to revoke Longrich's grant, a university spokesperson told *Nature*: "We respect this decision by the Leverhulme Trust and appreciate the fact they will continue to support the existing PhD students."

The spokesperson added: "All staff and students have a right to be treated, and have an obligation to treat others, with dignity and respect. The University has previously issued a statement about the result of a disciplinary hearing. We have been supporting students and staff throughout this period."

Nature has asked Longrich for comment on the revocation of his grant and the investigation conducted by his university, and is awaiting his response.

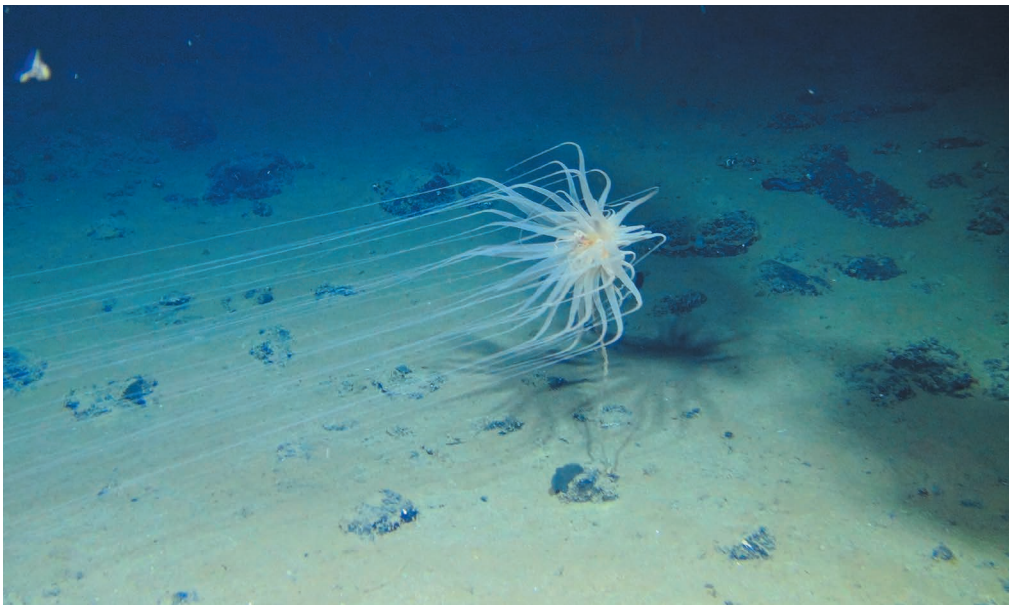
A STRING OF ALLEGATIONS

The news follows two other inquiries into bullying at prominent UK research institutions in the past month, one of which also led to a grant being revoked.

On 17 August, the Wellcome Trust, one of the world's largest research-funding charities, announced that it would revoke a £3.5-million ($4.5-million) grant from leading cancer geneticist Nazneen Rahman, after allegations that she had bullied scientists and other staff members when she worked at the Institute of Cancer Research (ICR) in London. The ICR concluded that there was enough evidence for some of the allegations to be considered at a disciplinary hearing, but Rahman resigned and the disciplinary hearing did not take place.

And at the end of August, the Wellcome Sanger Institute in Hinxton, UK, confirmed that it was investigating allegations of bullying there. ■

1. Martill, D. M., Tischlinger, H. & Longrich, N. R. *Science* **349**, 416–419 (2015).
2. Longrich, N. R. *J. Paleontol.* **84**, 681–694 (2010).



A new species of sea-anemone-like *Relicanthus* clings to a sponge stalk on Pacific Ocean floor.

D. J. AMON & C. R. SMITH

ECOLOGY

Deep-sea mining zone thrives with life

Discoveries come as nations prepare to mine sea bed.

BY AMY MAXMEN

Deep in the eastern central Pacific Ocean, on a stretch of sea floor nearly as big as the continental United States, researchers are discovering species faster than they can name them. And they are exploring newfound fossil beds of whales that lived up to 16 million years ago.

The findings — many reported for the first time last week at the Deep-Sea Biology Symposium in Monterey, California — have come as a shock. Some scientists had thought these vast underwater plains, 4,000–5,500 metres below the ocean surface, were relatively lifeless. But that is changing just as nations and corporations prepare to mine this patch of the Pacific sea bed for cobalt, manganese and other elements for use in technologies such as smartphones and electric cars.

Researchers are now pushing the International Seabed Authority (ISA), the body that oversees mining in international waters, to limit environmental damage from future activity. The ISA, which is developing rules for mining in the ocean, is accepting comments on a draft plan until 30 September. Its goal is to release final rules by 2020, clearing the way for mining to start.

"What we do right now is going to have huge implications for decades," says Diva Amon, a deep-sea biologist at the Natural History Museum in London. "We have a chance to do things as rigorously and responsibly as we can."

The ISA began issuing contracts to explore the Clarion–Clipperton Zone (CCZ), a 6-million-square-kilometre swathe of the Pacific Ocean floor that stretches from ▶

MORE ONLINE

TOP NEWS



First animal was a pancake-shaped prehistoric ocean dweller go.nature.com/2lea2jx

MORE NEWS

● Visa problems hamper African and Asian researchers go.nature.com/2oqxgcl
● Project aims to transform farming in poorest countries go.nature.com/2oeupxy
● Denisovan hybrid cave yields four more hominin bones go.nature.com/2r1w1cw

NATURE PODCAST

An ultra-thin, wearable biosensor, and a multi-shape metamaterial nature.com/nature/podcast

© 2018 Springer Nature Limited. All rights reserved.

Hawaii to Mexico, in 2001. The agency has given 29 companies permission to explore mining in international sea beds, including at 17 sites in the CCZ. Those companies, and the nations sponsoring them, must produce environmental assessments of their plots to satisfy the ISA's mandate to enable mining while preserving the ocean environment.

Sea-floor surveys are just beginning to map the vibrant life in areas marked for mining. Craig Smith, an oceanographer at the University of Hawaii at Manoa in Honolulu, helped lead expeditions in 2013 and 2015 to a territory in the eastern CCZ claimed by the United Kingdom. He was surprised to find hills and mountains rising across deep undersea plains, and life more diverse than that seen at similar depths elsewhere.

Seventy per cent of the 154 marine worm species found by Smith's team seem to be unknown to science. The researchers discovered sea cucumbers they call "gummy squirrels", and a new member of the order *Relicanthus* that is related to sea anemones. The species attaches itself to the stalk of a sponge and extends its tendrils into the current. The team also saw rare worms that resemble squid.

Adrian Glover, a deep-sea biologist at the Natural History Museum in London, has found that potato-sized nodules of manganese and other metals in the eastern CCZ harbour geometric sponges and other tiny, rarely seen invertebrates. And swathes of the sea floor are covered with enigmatic xenophyophores — ornate single-celled creatures that can be larger than a softball, and that exude slime as they feed. Most of the xenophyophores that scientists have found in the area were previously unknown, expanding the number of recognized species by about 30%.

But the wave of discoveries in the eastern CCZ is not limited to living species. Amon stunned the audience at the deep-sea symposium with photos of fossilized whale skulls encrusted in metallic residues. Her preliminary analyses suggest the bones belong to perhaps 6 extinct species of whale that died between 1 million and 16 million years ago. A study published in August suggests that modern beaked whales feed on the sea floor in the eastern CCZ (L. Marsh *et al. R. Soc. Open Sci.* **5**, 180286; 2018). The authors speculate that the whales ingest metallic nodules to regulate their buoyancy under water.

"What we do right now is going to have huge implications for decades."

Amon is among the scientists who argue that their findings should prompt the ISA to conserve a section of the eastern CCZ. In 2012, the agency set aside nine preserves in the CCZ, relying largely on satellite images showing the density of plankton in the sea. But none of these areas is in the east, where researchers have begun to document surprisingly complex sea-floor ecosystems.

And Smith is pushing the ISA to support research in the open waters above sea-floor mining zones. He and his colleagues say that silt and toxins discharged by mining could prevent some marine organisms from breathing and eating, block bioluminescent light that some use to attract prey and find mates, and pollute the food web.

Researchers will never know everything that lives on the deep-sea plains, so they must plan around this uncertainty as they advise the ISA, says Malcolm Clark, a marine scientist who sits on the agency's legal and technical commission. "Scientists have to recommend a course of action that includes a learning process, so that if things start to go off the rails we can get back on course," he says.

In the meantime, the ISA is under pressure to finalize its mining regulations by 2020 so that large-scale operations can begin. Japan began extracting deep-sea minerals late last year at a test site near the island of Okinawa. And a Belgian company, Global Sea Mineral Resources in Ostend, plans to test its equipment in the CCZ next year. ■

POLICY

NSF rolls out stringent anti-harassment policy

National Science Foundation mandates that institutions report infractions by grant holders.

BY ALEXANDRA WITZE

The US National Science Foundation (NSF) has unveiled its long-awaited policy to fight harassment by the scientists whose research it supports. Starting on 21 October, institutions that receive NSF grants must notify the agency of any finding related to harassment — including sexual harassment or sexual assault — by principal investigators (PIs) or co-PIs. Actions that must be reported include putting a scientist on leave during an investigation.

The rule will apply to all new grants and any extensions to existing grants made on or after that date. The NSF first proposed the policy in February, and received nearly 200 public comments on its initial draft.

It is the strictest action yet by a US research-funding agency on the topic of sexual harassment. On 17 September, the National Institutes of Health said that it would introduce a centralized system for reporting harassment, but it has not imposed any new conditions on the grants it hands out.

Experts say that the NSF policy is a good first step, but far from a final one. "This rule is wonderful news, and goes beyond what I expected," says Jane Willenbring, a geologist at the Scripps Institution of Oceanography in La Jolla, California, who has worked to improve safety at NSF-funded field sites, such as those in Antarctica. But she says that the NSF should consider investigating harassment complaints itself, rather than relying on the alleged perpetrator's academic institution to do the job.

Erika Marín-Spiotta, a geoscientist at the University of Wisconsin–Madison, says that she would have liked to have seen the NSF policy apply to current grants, not just those awarded or extended after 21 October. The new rule, she says, "is the bare minimum requirement" of what is needed.

NSF director France Córdova says the rule will not be the agency's final action against harassment. "NSF does not consider its work in trying to address harassment finished," she says.

MANDATORY DISCLOSURE

The rule will require an institution to notify the NSF within ten business days if it finds that a PI or co-PI on an NSF grant has committed harassment. Institutions are also required to report when they take any administrative action related to a harassment finding or investigation, such as putting a PI or co-PI on leave. The NSF policy does not, however, require institutions to notify the agency when they

© 2018 Springer Nature Limited. All rights reserved.

begin an investigation. "There's a delicate balance between receiving complaints and the due process people are entitled to," says Córdova. "An actual finding of determination is something concrete we can take action on."

Roughly 2,000 universities and other research institutions receive NSF funding. They are already legally responsible for complying with federal civil-rights laws — including the Title IX legislation that bars discrimination on the basis of sex, which has been widely used to battle sexual harassment and assault on campus. Many people do not report harassment through their home institutions, however, because they fear retaliation or assume that an investigation will not be thorough, Marín-Spiotta notes. The NSF's new policy allows people to report harassment directly to the agency in addition to, or rather than, going through an institution.

Since it first proposed the reporting rule in February, the NSF has received "five or six" notifications from institutions of findings involving harassment by grant recipients, and "probably at least twice that" number of notifications from individuals who have encountered or witnessed harassment involving grantees, says Robert Cosgrove, a compliance programme manager at the NSF.

Reports of harassment involving NSF grantees can be filed online at https://nsf.gov/harassment. Rhonda Davis, head of the NSF Office of Diversity and Inclusion, says she has added several staff members who are experienced in dealing with harassment to work on the reports.

Asked explicitly about bullying, Córdova noted that it is covered under "other forms of harassment" in the NSF rule.

BROADER ACTION

The push to deal with harassment in science continues to expand. On 15 September, the governing council of the American Association for the Advancement of Science in Washington DC voted to establish procedures to revoke fellowship honours for scientists found to have committed misconduct or ethical breaches, including harassment. And Congress has asked the US Government Accountability Office to open an investigation into how the various government funding agencies deal with sexual harassment involving their grantees.

Córdova says that the NSF and other agencies are waiting for the Senate to confirm a director for the White House Office of Science and Technology Policy to help coordinate anti-harassment efforts across agencies. Kelvin Droegemeier, a meteorologist at the University of Oklahoma in Norman, has been nominated for that post, but is not yet confirmed. ■



Thousands of years ago, some young children laboured in salt mines in what is now Austria.

NHM VIENNA/HANS RESCHREITER

ARCHAEOLOGY

Prehistoric children toiled at tough tasks

Kids as young as eight worked as brickmakers and miners.

BY TRACI WATSON

A surge of interest in the archaeology of childhood is revealing details of the skilled and sometimes back-breaking work that youngsters performed hundreds to thousands of years ago.

Their tasks included mining salt and forming bricks. Some children were already learning to create clay vessels by the time they were six years old. Researchers presented several of these findings at a meeting of the European Association of Archaeologists (EAA) in Barcelona, Spain, earlier this month.

Artefacts and skeletal remains that provide details of child labour from so long ago are still relatively sparse. But scholars are showing increasing interest in the subject, says archaeologist Mélie Le Roy at the Mediterranean Laboratory of Prehistory in Europe and Africa-UMR 7269 in Aix-en-Provence, France, who was one of the session organizers.

"In the next years," she says, "we will find more and more evidence that children were participating early in their lives in economic society."

WORKED TO THE BONE

Researchers paid little heed to children in the archaeological record until recently. But in the 1990s, more archaeologists began to examine the role of women in the past. That led some scientists to start studying other overlooked groups — including children.

Recent work suggests that some prehistoric youngsters toiled in harsh environments, including mines. Researchers excavating the ancient salt mines of Hallstatt, Austria, have discovered a child-sized leather cap dated to 1000–1300 BC, along with very small mining picks, says archaeologist Hans Reschreiter at the Natural History Museum of Vienna. This suggests that children were working in these mines at least two centuries earlier than scientists had thought.

To confirm this, Reschreiter and his colleagues plan to test heaps of human excrement found in the Bronze Age section of the ▶

© 2018 Springer Nature Limited. All rights reserved.

mine for sex hormones, which younger children would lack.

But ancient child labour wasn't always so back-breaking. When Le Roy analysed a jumble of skeletal remains from prehistoric tombs in France, she found three baby teeth with cylindrical grooves. Such abrasions form when people use their teeth for repeated, forceful stretching and softening of animal tendon or plant material, Le Roy says. The material was probably used for sewing or making baskets, she adds (M. Le Roy *Ardèche Archéol.* **35**, 12–18; 2018).

The teeth belonged to two children between the ages of one and nine. They date to 2100–3500 BC, making them some of the oldest evidence that children were engaged in skilled labour. Le Roy is about to start surveying human remains from more than 30 French burial sites from the same time period, and expects to find more evidence of young children at work.

LEAVING THEIR MARK

Other researchers are looking to artefacts rather than skeletons for information on child labour. When archaeologist Steven Dorland at the University of Toronto, Canada, examined ceramic shards from a prehistoric village in what is now southern Canada, he saw minuscule fingernail marks in the fifteenth-century debris. The size of the indentations showed that kids aged six or younger were forming clay vessels (S. G. H. Dorland *J. Archaeol. Sci. Rep.* **21**, 298–304; 2018).

In some modern communities, only pots of a certain quality make it to the kiln. But at Dorland's site, youngsters' misshapen starter pots were also fired. "It shows children in those societies had a certain level of social value," he says.

Even after the advent of written records — which can document the presence of youngsters in the workforce — archaeological evidence can provide powerful illumination of the role of children. Bricks and roof tiles excavated from a Lithuanian castle, dated to between the thirteenth and seventeenth centuries, still bear the fingerprints of their young creators.

Analysis of the prints' ridges suggests that children between the ages of 8 and 13 made more than 10% of the recovered building materials, said archaeologist Povilas Blaževičius at the National Museum of the Palace of the Grand Dukes of Lithuania in Vilnius during his EAA presentation.

Lithuania lacks written sources about children in the historical workforce, Blaževičius says. That leaves physical traces as the only evidence of their efforts there centuries ago. "When we have fingerprints of a child inside a pot, we definitely show that a child formed it," he says. "For me as an archaeologist, it's another way to find children in past societies." ■

SCHRÖDINGER'S CAT

Quantum puzzle baffles physicists

New twist on thought experiment yields conflicting results.

BY DAVIDE CASTELVECCHI

In the world's most famous thought experiment, physicist Erwin Schrödinger described how a cat in a box could be in an uncertain predicament. The peculiar rules of quantum theory meant that it could be both dead and alive, until the box was opened and the cat's state measured. Now, two physicists have devised a modern version of the paradox — with shocking implications — by replacing the cat with a physicist doing experiments.

Quantum theory has a long history of thought experiments, and most are used to point to weaknesses in various interpretations of quantum mechanics. But the latest version is unusual: it shows that if the standard interpretation of quantum mechanics is correct, then different experimenters can reach opposite conclusions about what the physicist in the box has measured. This means that quantum theory contradicts itself.

Physicists have debated this conceptual experiment with gusto for more than two years — and it has left most researchers stumped, even in a field accustomed to weird concepts. "I think this is a whole new level of weirdness," says Matthew Leifer, a theoretical physicist at Chapman University in Orange, California.

The authors, Daniela Frauchiger and Renato Renner of the Swiss Federal Institute of Technology in Zurich, first posted their argument online in April 2016, and published a paper on 18 September (D. Frauchiger and R. Renner *Nature Commun.* **9**, 3711; 2018). (Frauchiger has now left academia.)

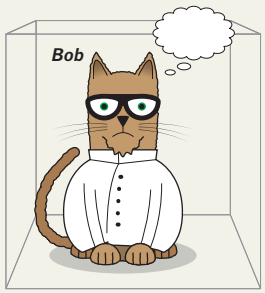
Quantum mechanics underlies nearly all of modern physics. But the answers it provides

NEW CATS IN TOWN

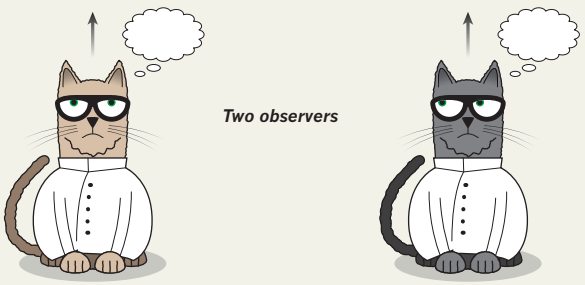
Physicists have devised a variation of the iconic Schrödinger's cat thought experiment that involves several players who understand quantum theory. But surprisingly, using the standard interpretation of quantum mechanics, the observers sometimes seem to come to different conclusions about a particular event — suggesting that the interpretation contradicts itself for complex systems.



Alice tosses a coin and, using her knowledge of quantum physics, sends a quantum message to Bob.



Using his knowledge of quantum theory, Bob can detect Alice's message and guess the result of her coin toss.



When the two observers open their boxes, in some situations they can conclude with certainty how the coin landed — but their conclusions are different. This means that the standard interpretation of quantum theory gives an inconsistent description of reality.

© 2018 Springer Nature Limited. All rights reserved.

can be frustratingly limited. Its equations cannot predict the exact outcome of a measurement — for example, of the position of an electron — only the probabilities that it can yield particular values.

Quantum objects such as electrons therefore live in a cloud of uncertainty, mathematically encoded in a 'wavefunction' that changes shape smoothly. But when a property such as an electron's position is measured, it always yields one precise value (and yields the same value again if measured immediately after).

The 'Copenhagen interpretation', formulated in the 1920s by Niels Bohr and Werner Heisenberg, is the most common way of understanding this. It says that the act of observing a quantum system makes the wavefunction 'collapse' from a spread-out curve to a single data point. But it leaves open the question of why different rules should apply to the quantum world of the atom and the classical world of laboratory measurements (and of everyday experience). But it is also reassuring: although quantum objects live in uncertain states, experimental observation happens in the classical realm and gives unambiguous results.

Now, Frauchiger and Renner are shaking physicists out of this comforting position. Their theoretical reasoning says that the basic Copenhagen picture — as well as other interpretations that share some of its basic assumptions — is not internally consistent.

Their scenario is more involved than Schrödinger's cat — proposed in 1935 — in which the feline lived in a box with a mechanism that would release a poison on the basis of a random occurrence, such as the decay of an atomic nucleus. In that case, the state of the cat was uncertain until the experimenter opened the box and checked it.

"I think this is a whole new level of weirdness."

In 1967, the physicist Eugene Wigner proposed a version of the paradox in which a physicist friend lived inside a box with a measuring device that could return one of two results, such as a coin showing heads or tails. Does the wavefunction collapse when Wigner's friend becomes aware of the result? One school of thought says that it does, suggesting that consciousness is outside the quantum realm. But if quantum mechanics applies to the physicist, then she should be in an uncertain state that combines both outcomes until Wigner opens the box.

Frauchiger and Renner have two Wigners, each doing an experiment on a physicist friend whom they keep in a box (see 'New cats in town'). One of the two friends (call her Alice) can toss a coin and — using her knowledge of quantum physics — prepare a quantum message to send to the other friend (call him Bob). Using his knowledge of quantum theory, Bob can detect Alice's message and guess the result of her coin toss. When the two Wigners open their boxes, in some situations they can conclude with certainty which side the coin landed on, Renner says — but occasionally their conclusions are inconsistent.

The result has triggered heated responses from quantum theorists. "Some get emotional," Renner says. And different researchers tend to draw different conclusions. "Most people claim that the experiment shows that their interpretation is the only one that is correct."

For Leifer, producing inconsistent results should not necessarily be a deal breaker. Some interpretations of quantum mechanics already allow for views of reality that depend on perspective. Robert Spekkens, a theoretical physicist at the Perimeter Institute for Theoretical Physics in Waterloo, Canada, says that the way out of the paradox could hide in some subtle assumptions in the argument, in particular in the communication between Alice and Bob.

"To my mind, there's a lot of situations where taking somebody's knowledge on board involves some translation of their knowledge," he says: perhaps the inconsistency arises from Bob not interpreting Alice's message properly. But he admits that he has not found a solution yet. ■

© 2018 Springer Nature Limited. All rights reserved.



THE BIG PUSH FOR CHINESE MEDICINE

For the first time, the World Health Organization will recognize traditional medicine in its influential global medical compendium.

BY DAVID CYRANOSKI

Choi Seung-hoon thought he had an impossible assignment. On a grey autumn day in Beijing in 2004, he embarked on a marathon effort to get a couple of dozen representatives from Asian nations to boil down thousands of years of knowledge about traditional Chinese medicine into one tidy classification system.

Because practices vary greatly by region, the doctors spent endless hours in meetings that dragged over years, debating the correct location of acupuncture points and less commonly known concepts such as 'triple energizer meridian' syndrome. There were numerous skirmishes between China, Japan, South Korea and other countries as they vied to get their favoured version of traditional Chinese medicine (TCM) included in the catalogue. "Each country was

A practitioner of traditional Chinese medicine treats a patient in Zhejiang province in China.

© 2018 Springer Nature Limited. All rights reserved.

CHINA DAILY/REUTERS

concerned how many terms or contents of its own would be selected," says Choi, then the adviser on traditional medicine for the Manila-based western Pacific office of the World Health Organization (WHO).

But over the next few years, they came to agree on a list of 3,106 terms and then adopted English translations — a key tool for expanding the reach of the practices.

And next year sees the crowning moment for Choi's committee, when the WHO's governing body, the World Health Assembly, adopts the 11th version of the organization's global compendium — known as the International Statistical Classification of Diseases and Related Health Problems (ICD). For the first time, the ICD will include details about traditional medicines.

The global reach of the reference source is unparalleled. The document categorizes thousands of diseases and diagnoses and sets the medical agenda in more than 100 countries. It influences how physicians make diagnoses, how insurance companies determine coverage, how epidemiologists ground their research and how health officials interpret mortality statistics.

The work of Choi's committee will be enshrined in Chapter 26, which will feature a classification system on traditional medicine. The impact is likely to be profound. Choi and others expect that the inclusion of TCM will speed up the already accelerating proliferation of the practices and eventually help them to become an integral part of global health care. "It will definitely change medicine around the world," says Choi, now the board chair of the National Development Institute of Korean Medicine in Gyeongsan.

Whether this is a good thing depends on whom you talk to. For Chinese leaders, the timing could not be better. Over the past few years, the country has been aggressively promoting TCM on the international stage both for expanding its global influence and for a share of the estimated US$50-billion global market.

Medical-tourism hotspots in China are drawing tens of thousands of foreigners for TCM. Overseas, China has opened TCM centres in more than two dozen cities, including Barcelona, Budapest and Dubai in the past three years, and pumped up sales of traditional remedies. And the WHO has been avidly supporting traditional medicines, above all TCM, as a step towards its long-term goal of universal health care. According to the agency, traditional treatments are less costly and more accessible than Western medicine in some countries.

Many Western-trained physicians and biomedical scientists are deeply concerned, however. Critics view TCM practices as unscientific, unsupported by clinical trials, and sometimes dangerous: China's drug regulator gets more than 230,000 reports of adverse effects from TCM each year.

With so many questions about TCM's effectiveness and safety, some experts wonder why the WHO is increasing support for such practices. One of them is Donald Marcus, an immunologist and professor emeritus at Baylor College of Medicine in Houston, Texas, and a prominent TCM critic. In his opinion, "at some point, everyone will ask: why is the WHO letting people get sick?"

DIFFERENT APPROACH

TCM is based on theories about qi, a vital energy, which is said to flow along channels called meridians and help the body to maintain health. In acupuncture, needles puncture the skin to tap into any of the hundreds of points on the meridians where the flow of qi can be redirected to restore health. Treatments, whether acupuncture or herbal remedies, are also said to work by rebalancing forces known as yin and yang.

Practitioners of TCM and Western-trained physicians have often eyed each other suspiciously. The Western convention is to seek well-defined, well-tested causes to explain a disease state. And it typically requires randomized, controlled clinical trials that provide statistical evidence that a drug works.

> "IT WILL DEFINITELY CHANGE MEDICINE AROUND THE WORLD."

From the TCM perspective, this is too simplistic. Factors that determine health are specific to individuals. Drawing conclusions from large groups is difficult, if not impossible. And the remedies are often a mix of a dozen or more ingredients with mechanisms that cannot, they say, be reduced to a single factor.

There has, however, been something of a détente. Organizations steeped in the Western conventions, such as the US National Institutes of Health (NIH), have created units to research traditional medicines and practices. And TCM practitioners are increasingly looking for proof of efficacy in clinical trials. They often speak of the need to modernize and standardize TCM.

Chapter 26 is meant to be a standard reference that all practitioners can use to help diagnose disease and assess possible causes. For example, 'wasting thirst syndrome' is characterized by excessive hunger and increased urination and explained by "factors which deplete yin fluids in the lung, spleen or kidney systems and generate fire and heat in the body". On the basis of those observations, physicians can work out how to treat them. The patient, who would probably be diagnosed as diabetic by a Western doctor, would probably be prescribed acupuncture, various tonics and moxibustion — in which practitioners burn herbs near the skin of the patient. Spinach tea, celery, soya beans and other 'cooling' foods would also be recommended.

TCM practitioners around the world are gearing up for Chapter 26, which is set to be implemented by WHO member states in 2022. "For the first time in history, ICD codes will include terminology such as Spleen Qi Deficiency or Liver Qi Stagnation," reads a post on the website of Five Branches University, a TCM training and research institution based in San Jose, California, which worked with the WHO on a field trial of the diagnostic criteria in Chapter 26.

Critics argue that there is no physiological evidence that qi or meridians exist, and scant evidence that TCM works. There have been just a handful of cases in which Chinese herbal treatments have proved effective in randomized controlled clinical trials. One notable product that has emerged from TCM is artemisinin. First isolated by Youyou Tu at the China Academy of Traditional Chinese Medicine in Beijing, the molecule is now a powerful treatment for malaria and won Tu the Nobel Prize in Physiology or Medicine in 2015.

But scientists have spent millions of dollars on randomized trials of other TCM medicines and therapies with little success. In one of the most comprehensive assessments, researchers at the University of Maryland school of medicine in Baltimore surveyed 70 systematic reviews measuring the effectiveness of traditional medicines, including acupuncture. None of those studies could reach a solid conclusion because the evidence was either too sparse or of poor quality[1]. The NIH's National Center for Complementary and Integrative Health in Bethesda, Maryland, concludes that "for most conditions, there is not enough rigorous scientific evidence to know whether TCM methods work for the conditions for which they are used".

In response to queries by *Nature,* the WHO said that its Traditional Medicine Strategy "provides guidance to Member States and other stakeholders for regulation and integration, of safe and quality assured traditional and complementary medicine products, practices, and practitioners". It emphasized that the goal of the strategy "is to promote the safe and effective use of traditional medicine by regulating, researching and integrating traditional medicine products, practitioners and practice into health systems, where appropriate".

China's support of TCM started with former leader Mao Zedong, who reportedly didn't believe in it but thought it a could reach underserved populations. Current Chinese President Xi Jinping has strongly supported TCM and, in 2016, the powerful state council developed a national strategy that promised universal access

© 2018 Springer Nature Limited. All rights reserved.

to the practices by 2020 and a booming industry by 2030. That strategy includes supporting TCM tourism, which steers large numbers of people to clinics in China. Every year, tens of thousands of mostly Russian tourists flock to Hainan off the southern coast seeking relief through TCM. The government has plans to build 15 TCM 'model zones' similar to the one in Hainan by 2020.

The country also has global ambitions. China's Belt and Road trade initiative calls for creating 30 centres by 2020 to provide TCM medical services and education, and to spread its influence. By the end of 2017, 17 centres had sprung up in countries such as the United Arab Emirates, Hungary, Kazakhstan and Malaysia.

The ties are paying off. Sales of TCM herbal medicines and other related products exported to Belt and Road countries surged by 54% between 2016 and 2017, to a total of US$295 million.



DAVID GRAY/REUTERS

A pharmacy in a traditional-medicine hospital in Beijing dispenses medications.

TIGHT TIES

The WHO's support applies to all traditional medicines, but its relationship with Chinese medicine, and with China, has grown especially close, in particular during the tenure of Margaret Chan, who ran the organization from 2006 to 2017. In Beijing in November 2016, Chan gave an address full of praise for China's advances in public health and its plans to spread traditional medicine. "What the country does well at home carries a distinctive prestige when exported elsewhere," she said.

Chan has supported traditional medicines, and specifically TCM, and has worked closely with China to promote this vision. In 2014, the WHO released a ten-year strategy that aims to integrate traditional medicines into modern medical care to achieve universal health coverage. The document calls on member states to develop health-care facilities for traditional medicine, to ensure that insurance companies and reimbursement systems consider supporting traditional medicines and to promote education in the practices.

In the same year, Chan wrote an introduction to a supplement that ran in *Science* and was sponsored by the Beijing University of Chinese Medicine and Hong Kong Baptist University[2]. (*Nature* ran a similar paid-for supplement in 2011.) Chan wrote that traditional medicines are "often seen as more accessible, more affordable, and more acceptable to people and can therefore also represent a tool to help achieve universal health coverage". In a 2016 speech in Singapore, Chan said that TCM has excelled at preventing or delaying heart disease because it "pioneered interventions like healthy and balanced diets, exercise, herbal remedies and ways to reduce everyday stress".

But many Western physicians and scientists doubt that the herbal remedies and various other components of TCM or other traditional medicines have much to offer in their current use. They grant that TCM herbs might turn up useful molecules (many Western drugs are derived from plants, after all), but worry that TCM could replace proven drugs or be potentially dangerous.

Arthur Grollman, a cancer researcher at Stony Brook University in New York, has published work showing how aristolochic acid, an ingredient in many TCM remedies, can cause kidney failure and cancer[3]. He thinks that WHO documents should pay more attention to the risks of remedies that contain the chemical, which are still widely used.

For some scientists, the WHO's embrace of TCM is perplexing. "I thought the WHO was committed to evidence-based medicine," says Richard Peto, a statistician and epidemiologist at the University of Oxford, UK.

Many physicians and researchers also find the WHO's declarations about traditional medicine hard to parse. Various WHO documents call for the integration of "traditional medicine, of proven quality, safety and efficacy". But the agency does not say which traditional medicines and diagnostics are proven. Wu Linlin, a WHO representative in the Beijing office, told *Nature* that the "WHO does not endorse particular traditional and complementary medicine procedures or remedies".

But that stands in sharp contrast to the WHO's actions in other areas. The agency gives member countries specific advice on what vaccines and drugs to use and what foods to avoid. With traditional medicines, however, the specifics are mostly omitted. The WHO website carries some warnings and states that aristolochic acid is a carcinogen. But with the repeated emphasis on integrating traditional medicine, the message is clear, says Marcus. In his view, "the WHO is clearly saying these are safe and effective medicines".

Nature tried to contact Chan multiple times through the WHO, but the agency says that she is not answering questions on matters related to the WHO.

MONEY MATTERS

Despite the concern over the WHO's decision to include TCM, even critics of the practices say that Chapter 26 could serve a constructive purpose. Peto says that Chapter 26 could help researchers to gather data on adverse reactions and what kinds of traditional treatments people are getting. "But if the aim is to endorse these things, it is inappropriate," he says.

For those steeped in Western medicine, the continued spread of traditional treatments is worrisome. TCM practitioners increasingly talk of replacing proven Western medicines with traditional substitutes, where there is a cost advantage. Grollman thinks that ICD-11 is heading in that direction. Seventy per cent of money spent on health care globally is reimbursed or allocated on the basis of ICD information. Now TCM will be part of that system.

"The thing they want is to make it sound official and be recognized by the insurance companies. Because it's relatively low cost, insurance companies will accept it," says Grollman.

Many others agree that the WHO's decision will help the spread of TCM. Inclusion in ICD-11 is "a powerful tool for [health-care] providers to say this is legitimate medicine" to insurers, says Ryan Abbott, a medical doctor who has also trained in TCM and is a faculty member at the University of California, Los Angeles, Center for East–West Medicine. The WHO's action regarding TCM, he says, "is a mainstream acceptance that will have significant impact around the world". ■

David Cyranoski *reports for* Nature *from Shanghai.*

1. Manheimer, E. *et al. J. Altern. Complement. Med.* **15**, 1001–1014 (2009).
2. Chan, M. *Science* **346**, S2 (2014).
3. Grollman, A. P. *et al. Adv. Mol. Toxicol.* **3**, 211–227 (2009).

© 2018 Springer Nature Limited. All rights reserved.

SIX MONTHS UNTIL BREXIT:

HOW SCIENTISTS ARE PREPARING FOR THE SPLIT

BY EHSAN MASOOD, INGA VESPER AND RICHARD VAN NOORDEN

The incredible is fast becoming the inevitable. Just six months from now, unless a political earthquake intervenes, the United Kingdom must leave the European Union. Yet the details of the March 2019 break-up are still muddy, and scientists are growing increasingly anxious — and angry — about how Brexit will alter their research and their lives.

The UK government has made plenty of promises to researchers: for instance, it says it is prepared to pay into EU science programmes after Brexit so that British scientists can apply for funding. And it has guaranteed that it will support existing EU research grants up to 2020 that might be jeopardized by the split. (The EU, for its part, has hinted that it would like to open up its next research programme to greater participation for non-EU countries.) Politicians have also assured non-UK academics from the EU — who make up one-sixth of British university staff (see 'Brexit risks') — that they need not fear losing their rights to work or claims to health benefits, although many are hurriedly applying for expensive residency permits to ward off potential disaster.

But these promises might mean little if — as seems plausible — Britain and the EU fail to agree on the terms of the departure. If a no-deal split happens, for instance, British science would instantly lose access to at least three of the major funding streams under the EU's Horizon 2020 funding programme. These streams alone have provided around €2 billion of the €4.8 billion (US$5.5 billion) that the United Kingdom has won from Horizon 2020 since 2014. And in a no-deal future, imports and exports of essential goods — including food, scientific equipment and medicines — would probably be disrupted.

It seems likely that, even if terms are agreed, the free-and-easy flow of researchers between Britain and the EU will stop: last week, an influential committee set up to inform UK migration policy said that European workers should be subject to the same visa rules as other migrants, although Britain might choose to offer preferential treatment as part of trade negotiations.

The 'nothing is agreed until everything is agreed' nature of Brexit means that uncertainty still reigns. Researchers are trying to prepare, but there are signs that the United Kingdom might already be losing out on EU research funds (see 'Waning influence?'). *Nature* presents the views of five scientists and campaigners on the stakes for science ahead of the Brexit crunch — including one who is fighting for a second vote to stop it all.

"The uncertainty is already stunting science."

Edith Heard, epigeneticist, Curie Institute, Paris

Edith Heard was devastated when she heard news of the Brexit referendum while at her home in France. "I didn't sleep that night," says the British scientist, who has lived across the channel since 1990. Heard directs the genetics and developmental biology department at the Curie Institute in Paris and is a fellow of London's Royal Society, but she was not allowed to vote on the Brexit measure because she had lived abroad for more than 15 years.

Now, like many other UK scientists on the continent who are unsure of their future, she is taking on dual nationality. She is applying for French citizenship. "The British scientists I talk to here have all either taken on European nationalities or are thinking of doing it if they can — because no one really knows what the consequences of Brexit will be," she says.

The view from Europe, says Heard, is that the uncertainty is already affecting research. "I've noticed that since the vote, UK scientists can't play as prominent a role in European projects. They can be part of them, but have drawn back from leadership roles because of the risk that proposals could be compromised in the future, when it's not clear what's going to happen," she says. "That's stunting science."

Brexit has already influenced Heard's scientific life. After the vote, she was approached to apply to be the next director of the European Molecular Biology Laboratory (EMBL) — a pan-European organization with six sites across the continent, including the European Bioinformatics Institute in Hinxton, UK. "I was so angry and distraught about what Brexit meant for European science that I saw it as my duty to do the job," she says. Heard will start as director next January, moving her laboratory — and family — to an EMBL site in Heidelberg, Germany. The United Kingdom's membership of EMBL will not change after Brexit because it was agreed independently of the EU. Heard sees EMBL's Hinxton site as "a little island of Europe within the UK", and a chance to build bridges if Brexit happens.

French researchers, too, are shocked by the situation and preparing for the split, she says. Scientists are trying to secure funds to maintain UK–EU working partnerships, despite the fogginess over future rules on collaboration and whether scientists will be able to move freely between the United Kingdom and Europe. "There's a will to move on and make sure science can happen," she says. "But at the moment, everyone is in this grey area of the unknown."

© 2018 Springer Nature Limited. All rights reserved.



Italian archaeologist Chiara Bonacchi has studied how perceptions of Britain's Roman past are shaping views on Brexit today.

KIERAN DODDS FOR *NATURE*

"I have no intention of leaving."

Chiara Bonacchi, archaeologist, Stirling University, UK

Brexit has injected a large dose of uncertainty into the lives of European researchers working in Britain. But archaeologist Chiara Bonacchi, at the University of Stirling, UK, is clear about her future. "I have no intention of leaving," she says emphatically.

Bonacchi, an Italian citizen who obtained British residency in June and is now applying for citizenship, says she knows of academics who have been lured away from other UK universities, following the Brexit vote, to professorships in Germany and Austria. But for her, the benefits of staying outweigh future risks. "I chose the UK because it gave me the opportunity to develop as an independent researcher," she says. By contrast, she says, academics in Italy have "very little space to grow" — with fewer opportunities to build their own teams and frame their own research questions until later in their careers.

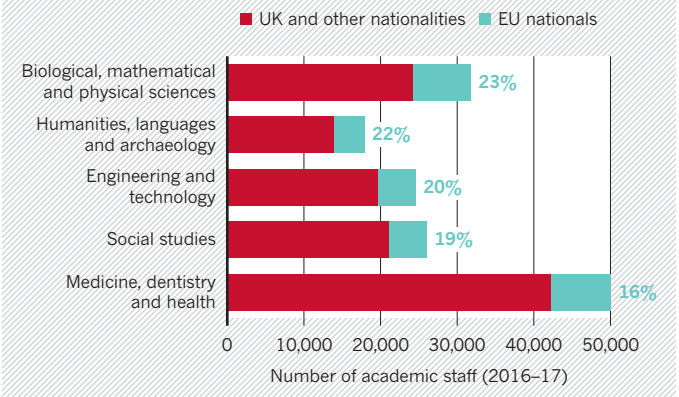
Bonacchi has even found a way to meld Brexit with her studies of how materials and cultural ideas derived from Britain's Iron Age and Roman past are drawn on today. Together with colleagues at University College London, she searched Facebook posts on Brexit for keywords relating to this historical period. "Roman" — as in the Roman empire — was one of the most commonly used of such terms, she says, found in more than 2,500 posts and comments (C. Bonacchi *et al. J. Soc. Archaeol.* **18**, 174–192; 2018).

Some Facebook posters who wanted Britain to remain in the European Union invoked the Roman empire as an early example of a civilizing power that brought benefits to Britain. But others, who backed Brexit, emphasized British resistance to Roman violence and oppression, and they characterized the Roman empire as a dictatorship.

These viewpoints match with wider discussions about whether people perceive Roman Britain as the origin of European civilization or as an enemy occupation — and illustrate how perceptions of Britain's past are still colouring its politics and society today. "I want to understand how the ways in which people perceive the past shape their political views and the kind of future they want to build," says Bonacchi.

BREXIT RISKS

Non-British EU staff make up an average of 17% of academics at UK universities — but the proportion is higher in the sciences.



SOURCE: HESA

© 2018 Springer Nature Limited. All rights reserved.

"If it weren't for Brexit, these people would have joined my lab."

Cesare Terracciano, cardiologist, Imperial College London

In 1991, Cesare Terracciano, a freshly graduated cardiologist from Rome, won a European fellowship to study heart disease. He grabbed his Italian passport and moved to the United Kingdom, renowned for its excellence and scientific collaborations. Today, he is a principal investigator at one of the world's leading research centres looking at how to repair damaged hearts, the British Heart Foundation's Cardiovascular Regenerative Centre in London.

But since the 2016 vote for Brexit, the centre has struggled to attract the best talent. Put off by uncertainty about their future in Britain, Cesare says, up-and-coming European cardiologists might not now make the same decision that he did.

Terracciano's lab is strongly reliant on overseas talent: of 13 staff members (including himself), 10 are non-UK Europeans. But in December 2016 and December 2017, two European postdocs whom Terracciano tried to recruit said they weren't coming. "They changed their minds because they didn't like the uncertainty of whether they can stay in the United Kingdom to develop their career," he says. "They told me: if there was no Brexit, we would definitely join your lab, but now we don't want to do that." He did find other candidates, one from Greece and one from Britain.

Terracciano is well funded from UK sources, so he is unlikely to face immediate financial consequences. Nor does he want to leave the United Kingdom: his wife and three children are British. But his life has been affected in other ways — he has a speech impairment, and the specialized therapist he has worked with for years has decided to move back to Ireland because of Brexit.

Terracciano's current team is extremely worried, too. "Our staff are very nervous about Brexit, because they do not know what is going to happen to them. We know nothing and we cannot give them reassurances," he says.

"Research money is already flowing away from the United Kingdom."

Silvana Muscella, chief executive, Trust-IT Services, UK

"I hope I will just wake up one day and find Brexit is not going to happen," says Silvana Muscella, a London-born entrepreneur. "All I want is to be able to go to my partners and say: 'Listen, guys, we made a stupid mistake. Let's go back to the drawing board and sort this all out.'"

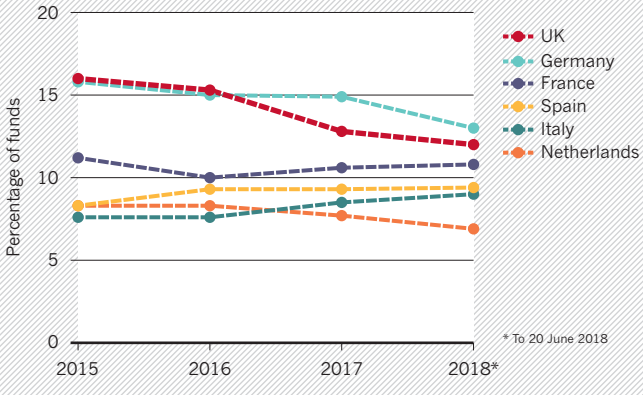
Muscella's software-services and marketing firm Trust-IT Services, in Enfield, UK, has won more than £15 million (US$19.7 million) from EU research funders over the past decade. It helps to coordinate projects such as the European Open Science Cloud, a repository that provides access to publicly funded data and research. But after the Brexit vote, Muscella largely gave up bidding for EU research money through the firm. Instead, she has switched to applying through a small company in Pisa, Italy, called COMMpla, which she set up in 2010 to serve commercial clients there. She is using the Italian company because potential collaborators told her that although they wanted Trust-IT on their projects, they would prefer a European firm because of uncertainty about the status of future British participation.

"These days of uncertainty provoke a sort of stigma attached to UK companies when people put a consortium together," Muscella says. "When a consortium is too UK-heavy, it is changed."

It's not clear what will happen to Trust-IT's existing projects after March 2019. If the United Kingdom is unable to participate in collaborative programmes after that point, Muscella will try — but might not be allowed — to transfer her funds to COMMpla. But even if UK politicians strike a deal to retain access to EU research programmes, the Brexit vote will still have had an impact, Muscella says. "Psychologically, it has changed how our partners approach us. The trust has completely gone," she says. "This will take a generation to heal."

WANING INFLUENCE?

Britain's annual share of funding from the EU's Horizon 2020 research programme is on the decline.



SOURCE: HORIZON 2020 DASHBOARD

"We need to stop Brexit."

Mike Galsworthy, co-founder, Scientists for EU, London

Ever since the Brexit vote, the campaign group Scientists for EU has been fighting for politicians to take scientists' interests into account (*Nature* **543**, 600–601; 2017).

But with six months to go until Britain legally splits with the EU, there is still no clarity about what will happen next. Now, says Mike Galsworthy, a former research-policy analyst who co-founded the group, the campaign has changed direction — to try to stop Brexit happening at all.

This year, Scientists for EU, which is already influential among anti-Brexit campaigners, has joined forces with a national movement called People's Vote, which advocates a second referendum on Britain's EU membership. The movement also has the backing of a group called Healthier in the EU (in the process of being renamed NHS Against Brexit). This group was co-founded by Scientists for EU and represents the interests of EU nationals working in Britain's National Health Service.

The campaign shifted its focus, says Galsworthy, because he thinks too little time remains to negotiate what he terms a "sensible" Brexit: a phased withdrawal of the United Kingdom from EU institutions over several years, rather than the instant break that politicians seem to be heading for. "We will either crash out, or ditch Brexit. The middle ground is eroding fast," he says.

Scientists for EU now has 150 groups of affiliated local volunteers across Britain who are knocking on doors to persuade people to back a second vote, Galsworthy says. And he thinks that if such a referendum happens, the result is likely to swing against Brexit: opinion polls now consistently favour Britain remaining in the EU, he points out. "If a majority of the public decided they would rather cancel Brexit, they fully have the right to do so," he says. ■ SEE GO.NATURE.COM/2RCLPHD FOR MORE PROFILES AND ANALYSIS

© 2018 Springer Nature Limited. All rights reserved.

COMMENT

DISASTERS Climate, fires and floods are linked — study them together **p.458**

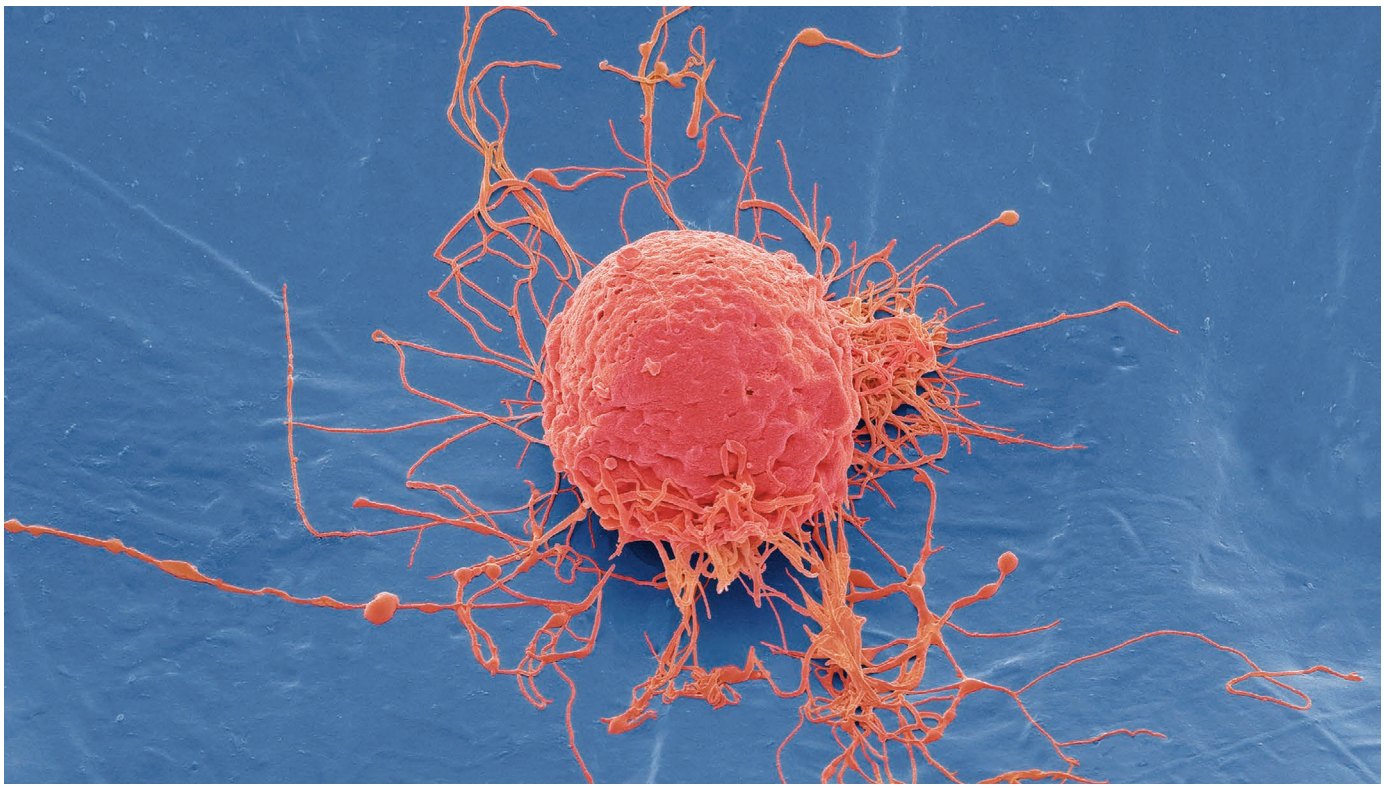


GENOMES Don't use genomics to excuse social inequality **p.461**

AUTHORSHIP Follow the film industry and list contributions instead **p.464**

CONFERENCES Do boring speakers talk for longer, or does it just feel that way? **p.464**

STEVE GSCHMEISSNER/SPL



Scanning electron micrograph of what is called a human mesenchymal stem cell, which some say can develop into bone, cartilage or fat cells.

Clear up this stem-cell mess

Confusion about mesenchymal stem cells is making it easier for people to sell unproven treatments, warn **Douglas Sipp**, **Pamela G. Robey** and **Leigh Turner**.

Various populations of cells in the adult human body have been the subject of controversy since the early 2000s. Contradictory findings about these haphazardly termed 'mesenchymal stem cells', including their origins, developmental potential, biological functions and possible therapeutic uses, have prompted biologists, clinicians and scientific societies to recommend that the term be revised or abandoned. Last year, even the author of the paper that first used the term mesenchymal stem cells (MSCs) called for a name change[1].

Tissue-specific stem cells, which have a limited ability to turn into other cell types, are the norm in most of the adult body. Several studies indicate that the variety of cells currently dropped into the MSC bucket will turn out to be various tissue-specific cell types, including stem cells[2].

Yet the name persists despite the evidence pointing to this, and almost two decades after questions about the validity of MSCs were first raised. A literature search indicates that, over the past 5 years, more than 3,000 research articles referring to MSCs have been published every year (see 'Tenacious term'). And several national regulatory agencies have now licensed MSC-based drugs, although most of these approvals have been provisional or are based on underpowered studies (see 'Doubtful drugs').

In our view, the wildly varying reports have helped MSCs to acquire a near-magical, all-things-to-all-people quality in the media and in the public mind[3] — hype that has been easy to exploit. MSCs have become the go-to cell type for many unproven stem-cell interventions. The confusion must be cleared up.

What is needed is a coordinated global effort to improve understanding of the biology of the cells currently termed MSCs, and a commitment from researchers, journal editors and others to use more-precise labels. We must develop standardized analyses of gene expression, including on a cell-by-cell basis, and rigorous assays to establish the precise products of cell differentiation in various tissues. Such ▶

© 2018 Springer Nature Limited. All rights reserved.

efforts could put an end to lingering questions about MSC identity and function, once and for all.

A CONTROVERSIAL CELL

The MSC concept dates to a 1991 paper[4] in which US biologist Arnold Caplan described the isolation of a type of stem cell found in bone marrow, building on reports by other groups (see, for example, ref. 5). Collectively, this work showed that certain cells in the supportive tissue of bone marrow (the stroma) could differentiate into cartilage, bone and fat, and provide some of the signals needed for haematopoietic stem cells (the immature precursors of all blood cells) to give rise to blood-cell types.

Over the next decade, there was an explosion in the number of tissue types in which MSCs were reported. And there was a similar jump in the diversity of cell types that MSCs were reportedly able to differentiate into. Yet by the early 2000s, it had become clear that separate labs were using different cell-surface markers to characterize MSCs. In 2006, a working group of the International Society for Cellular Therapy (ISCT) acknowledged the "inconsistencies and ambiguities" and recommended a new designation: multipotent mesenchymal stromal cells[6].

Despite the ISCT's recommendation, the stem-cell designation has proved strangely durable. Assays commonly used to evaluate a cell's 'stemness' are prone to misinterpretation. Many researchers have failed to take into account the developmental origins of the tissues they cite as sources of MSCs, or to observe the rigorous definition of a stem cell. Also, the very commonness of the MSC term seems to have consolidated its acceptance.

In the past few years, however, there has been further questioning of the MSC as a valid biological entity. A 2014 study[7] by researchers at the US Food and Drug Administration found almost no agreement in the molecular characterization of MSCs among the investigational new drug applications submitted to the agency. A 2016 study[2] likewise found that various cell populations from different tissues, all classed as MSCs, actually differ drastically in their gene expression and in their ability to differentiate. And last year, Caplan revealed that he no longer believes that MSCs are stem cells[1].

In that article, Caplan implores the scientific community to adopt yet another moniker: medicinal signalling cells. (According to Caplan, medicinal signalling cells, identified on the basis of cell-surface proteins and their ability to turn into other cell types *in vitro*, home in on sites of injury. There they secrete cocktails of proteins that modulate the immune response, reduce inflammation, promote wound healing and inhibit cell death.)

Since 1991, more than 32,000 Medline-indexed articles referring to "mesenchymal stem cells" have been published. Our reading of a subset of these — many hundreds over the past decade — suggests that the field continues to be a mess.

According to the literature, MSCs are most often isolated from bone marrow and adipose tissue (fat). They have also been identified in perinatal tissues — umbilical cord, Wharton's jelly (a gelatinous substance found in the umbilical cord), the amnion (the membrane surrounding the embryo) and the placenta — as well as in other sources, including baby teeth and menstrual blood. Some groups report that MSCs are most common in fat. Others state that they are associated with blood vessels throughout the body's connective tissues. Some note that MSCs are exceedingly rare; others say they are abundant.

Canonically, MSCs (including those from non-skeletal sources) are supposed to give rise to cartilage, bone and fat. But reports exist of their differentiation into muscle, endothelium and various cells of the heart, liver and kidney, as well as of the nervous and reproductive systems. Some researchers and clinical providers have even described MSCs as nearly pluripotent, meaning that they can differentiate into almost every cell type in the adult body. Moreover, there are countless claims of therapeutic uses for these cells in multiple unrelated diseases, ranging from arthritis and diabetes to Parkinson's disease and autism.

SALES PITCH

All of the disagreement in the scientific community about the nature of MSCs, and even about their existence, is almost certainly making it easier for businesses to sell treatments allegedly based on MSCs.

Direct-to-consumer marketing of unapproved stem-cell treatments for medical conditions has exploded in the past five years, particularly in the United States, Australia and Japan. A 2016 report[8] documented 351 US companies selling putative stem-cell treatments direct to consumers; almost half refer to MSCs in their marketing materials.

Businesses have been quick to capitalize on the conflicting claims in the literature[9]. To convey that MSCs can treat a wide range of diseases and injuries, firms selectively highlight those articles suggesting that MSCs are easy to harvest and can give rise to other cell types, or those indicating that the cells secrete all sorts of healing factors. (The large body of research indicating that tissue-specific stem cells have more limited roles is overlooked.)

Interestingly, some publicly traded companies have shown increasing reluctance to define their products as mesenchymal stem cells. Cytori Therapeutics in San Diego, California, a developer of devices for harvesting and processing cells from patients, has begun to describe its target cell population as 'adipose-derived regenerative cells' instead of calling them MSCs.

SOLUTIONS IN SCIENCE

What can be done to clear up the confusion? In our view, re-categorizing MSCs as 'stromal' or 'signalling' cells won't help.

Some researchers are following the ISCT recommendations, but judging by the thousands of papers on MSCs still being published every year, many are not. Meanwhile, Caplan's proposal could introduce more problems than it solves. There is strong evidence for the existence of a tissue-specific stem cell in bone-marrow stroma at least, albeit one with a limited ability to differentiate into other cell types. Also, the use of 'medicinal' instead of 'mesenchymal' could encourage the assumption that MSCs have broad therapeutic usefulness before robust evidence for this has been obtained.

We think that answers will be found in better science. After all, many researchers now doubt that a single multipotent stem cell (meaning one that can give rise to several other kinds of cell) is present throughout the adult body, thanks to the rigour and persistence of some stem-cell biologists.

The reliable identification of tissue-specific cells — stem cells or otherwise — should involve omics approaches, such as those designed to analyse the gene-expression patterns of a cell or its protein content, and rigorous assays to establish what a particular cell can differentiate into, either *in vivo*

DOUBTFUL DRUGS

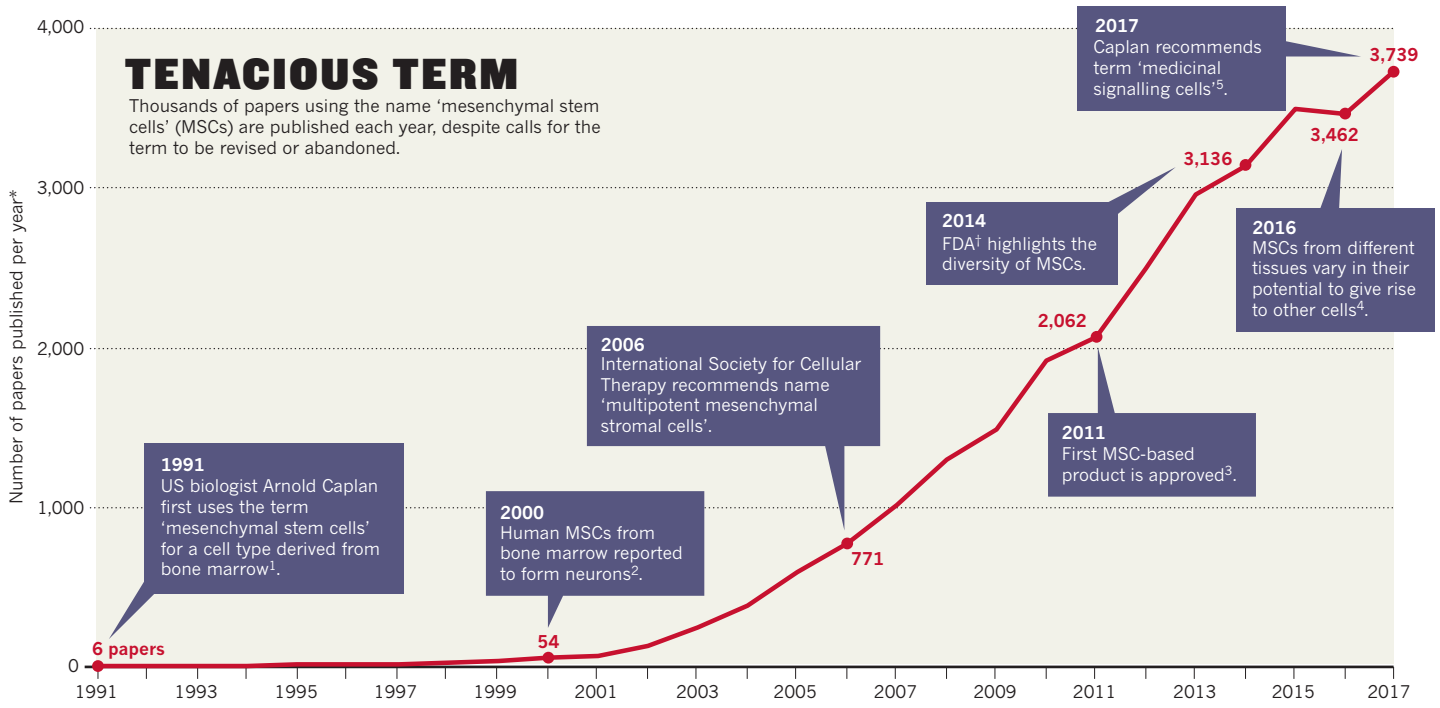
Clinical trials with MSCs fail to deliver

A move towards translational studies requires a robust understanding of the actual biological properties of the cell types currently called mesenchymal stem cells (MSCs).

So far, both the results and the quality of MSC-based clinical trials have been underwhelming. Take a 2014 meta-analysis of 49 trials using 'bone-marrow stem cells' (in many cases, 'bone-marrow MSCs') to treat cardiovascular disease, for instance. According to that analysis, the studies that scored better in terms of rigour were more likely to report less efficacy for MSC treatments than were those judged to be less rigorous[11].

Clinical studies using MSCs (or any stem cells) must adhere to the same standards of research design and oversight that apply to any responsible clinical trial before the cells are administered to human participants. **D.S., P.G.R. & L.T.**

© 2018 Springer Nature Limited. All rights reserved.



1. Caplan, A. I. *J. Orthop. Res.* **9**, 641–650 (1991). 2. Woodbury, D. *et al. J. Neurosci. Res.* **61**, 364–370 (2000). 3. Yang, H. *Nature Biotechnol.* **29**, 857 (2011).
4. Sacchetti, B. *et al. Stem Cell Rep.* **6**, 897–913 (2016). 5. Caplan, A. I. *Stem Cells Transl. Med.* **6**, 1445–1451 (2017).
*From PubMed search for "mesenchymal stem cells". †US Food and Drug Administration.

or *in vitro*. (Historically, inconsistencies in surface markers and stains, or the use of unreliable morphological characteristics, have led to frequent errors when identifying cell types[10].)

Ongoing initiatives such as the Human Cell Atlas, an effort to characterize all cells in the body, could shed light on the cellular components of specific tissues. But, ultimately, resolving the MSC identity crisis is likely to require a focused undertaking by stem-cell biologists — similar to a series of projects conducted in the 2000s by the International Stem Cell Forum to characterize pluripotent stem cells. (The forum is a collaboration designed to support stem-cell research.)

As part of the push for better science, regulatory agencies and editors of influential stem-cell and general scientific journals will need to develop and enforce rigorous methodological standards.

DISPEL THE MSC MYTH

Meanwhile, the community needs to stop subsuming multiple cell types under one catch-all phrase.

Scientific societies, such as the International Society for Stem Cell Research (ISSCR) and national stem-cell organizations, should consider whether MSC studies ought to be presented at stem-cell conferences under the MSC umbrella. Clinical-trial registries, such as clinicaltrials.gov, should also exercise heightened scrutiny over studies in which MSCs are described as the investigational product. (Hundreds of such studies are already listed on international clinical-research databases.) Groups studying tissue-specific stem cells should be encouraged to rethink their choice of terms. Both bone marrow and fat, for example, contain stem-cell populations that help to maintain these tissues. And referring to these as skeletal or adipose stem cells could help to vanquish the myth of a near-ubiquitous, all-powerful adult MSC[2]. Certainly, studies of poorly characterized cellular miscellanies should not be accepted without rigorous evidence of multipotency and self-renewal.

"In most cases, rigorous preclinical studies of these cells are limited or non-existent."

Journal editors, editorial boards and reviewers should similarly be more exacting when accepting or approving MSC reports for publication in journals. And funding bodies must consider whether studies using MSCs qualify for funds earmarked for stem-cell research.

Journalists can also play an important part in combating MSC misconceptions. When reporters write stories about clinics selling 'mesenchymal stem-cell treatments' and other purported stem-cell therapies, they should inform their audiences that scientists debate whether such cells are, in fact, stem cells. They should also make it clear that, in most cases, rigorous preclinical studies of these cells are limited or non-existent.

Meanwhile, organizations such as the ISSCR, the ISCT and national stem-cell societies should ensure that guidebooks, educational videos and other communication tools designed for people receiving 'adult stem-cell therapies' are updated to reflect the demise of the MSC as a viable concept.

Whatever the ultimate identity and biological activity of the cells formerly known as MSCs, the scientific principles that guide their clinical development and use must be the same as those for other new therapies: precision, validation, characterization, objectivity and the abandonment of terms and conceptual models that mislead more than they illuminate. ■

Douglas Sipp *is a researcher at RIKEN and project professor at Keio University School of Medicine and Global Research Institute, Tokyo, Japan.* **Pamela G. Robey** *is a senior investigator at the National Institute of Dental and Craniofacial Research, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA.* **Leigh Turner** *is associate professor at the Centre for Bioethics, University of Minnesota, Minneapolis, USA.*
e-mail: sipp@cdb.riken.jp

1. Caplan, A. I. *Stem Cells Transl. Med.* **6**, 1445–1451 (2017).
2. Sacchetti, B. *et al. Stem Cell Rep.* **6**, 897–913 (2016).
3. Caulfield, T., Sipp, D., Murry, C. E., Daley, G. Q. & Kimmelman, J. *Science* **352**, 776–777 (2016).
4. Caplan, A. I. *J. Orthop. Res.* **9**, 641–650 (1991).
5. Owen, M. & Friedenstein, A. J. *Ciba Found. Symp.* **136**, 42–60 (1988).
6. Dominici, M. *et al. Cytotherapy* **8**, 315–317 (2006).
7. Mendicino, M., Bailey, A. M., Wonnacott, K., Puri, R. K. & Bauer, S. R. *Cell Stem Cell* **14**, 141–145 (2014).
8. Turner, L. & Knoepfler, P. *Cell Stem Cell* **19**, 154–157 (2016).
9. Sipp, D. *et al. Sci. Transl. Med.* **9**, eaag0426 (2017).
10. Bianco, P., Robey, P. G. & Simmons, P. J. *Cell Stem Cell* **2**, 313–319 (2008).
11. Nowbar, A. N. *et al. Br. Med. J.* **348**, g2688 (2014).

L.T. declares competing non-financial interests; see go.nature.com/2pjhdai for details.

© 2018 Springer Nature Limited. All rights reserved.



A wildfire in Montecito, California, in December 2017. The following month, heavy rain falling on the burnt slopes caused a mudslide that killed 21 people.

ROBYN BECK/AFP/GETTY

How do natural hazards cascade to cause disasters?

Track connections between hurricanes, wildfires, climate change and other risks, urge **Amir AghaKouchak** and colleagues.

This has been an exceptional year so far for natural disasters. Typhoons in Asia and Hurricane Florence hitting the US east coast have caused extensive damage, flooding and mudslides. In the past two months, Scandinavia, Spain and Portugal, the United Kingdom, North America and South Africa experienced fierce forest blazes. Just outside Athens in July, one of the deadliest conflagrations in recorded history raged through coastal towns, killing 99 people. In the same month in Mendocino, California, an area larger than Los Angeles was scorched — more than 1,800 square kilometres — killing a firefighter and destroying almost 300 homes. Sweden suffered more than 50 wildfires, some even within the Arctic Circle.

The unprecedented severity of many of these fires might be a sign of global warming, with worse to come. But it also highlights how extreme events are connected. Many of the fires followed long periods of drought and record temperatures. Their occurrence also loads the dice for devastating future hazards.

Charred landscapes are more vulnerable to flooding and landslides. In January, a mudslide killed 21 people and injured more than 160 near Montecito, California. The month before, a wildfire had destroyed vegetation and destabilized the soil on the town's steep slopes. When a storm brought heavy rains, a 5-metre-high wave of mud, boulders and branches, travelling at 30 kilometres per hour, swept into people's homes.

Chains of adverse events such as these, cascading like toppling dominoes, will become more common as the world warms. Yet the scales of the cascading risks are still unknown. Risk analysts estimate the likelihood of single events, predicting rising frequencies of droughts, hurricanes and so on. They do not consider the web of connections between them. For example, sea-level rise will enhance coastal erosion and expose communities, infrastructure and ecosystems to damage from storms and surges.

Risk assessments should be expanded to consider cascading hazards. Otherwise, we cannot plan for the scale and nature of upcoming disasters. Researchers must find answers to these questions: how will climate change alter the risk of disastrous domino effects? What are the implications for the built environment? And what mitigation and adaptation measures are needed to cope

© 2018 Springer Nature Limited. All rights reserved.

with more severe interlinked disasters?

Here we outline how such a risk framework should be developed.

COMPOUND DISASTERS

The first step is for researchers and risk managers to recognize that climate change impacts do not occur in isolation, but are strongly coupled. For instance, droughts and heatwaves often occur together. Droughts lead to dry soils, which prevent solar energy from being released as evaporation, causing surface warming[1]. Across the United States, week-long heatwaves that coincide with periods of drought now occur twice as often as they did in the 1960s and 1970s (ref. 2).

Dry and warm conditions increase the risk of wildfires, which damage the soil and set the stage for later landslides and flooding. Snow and ice melt earlier, altering the timing of run-off. This has extended the fire season by 20% around the globe since the 1980s (refs 3,4). With less snow and ice in the Northern Hemisphere, the cooling effect provided by the reflection of sunlight from Earth's surface dropped by 10–20% between 1979 and 2008 (relative to its mean value during this period)[5].

And these links now spread further: wildfires are occurring at ever-higher elevations and latitudes (see 'More fires, more snowmelt'), where they remove the forest canopy and alter where and how snow accumulates. Soot deposited on the snow absorbs heat and speeds up melting. Likewise, dust released during droughts hastens melting, as has happened in the Upper Colorado River Basin[6]. Dust transported from arid regions of Africa influences the snow caps of Europe, North America and Asia.

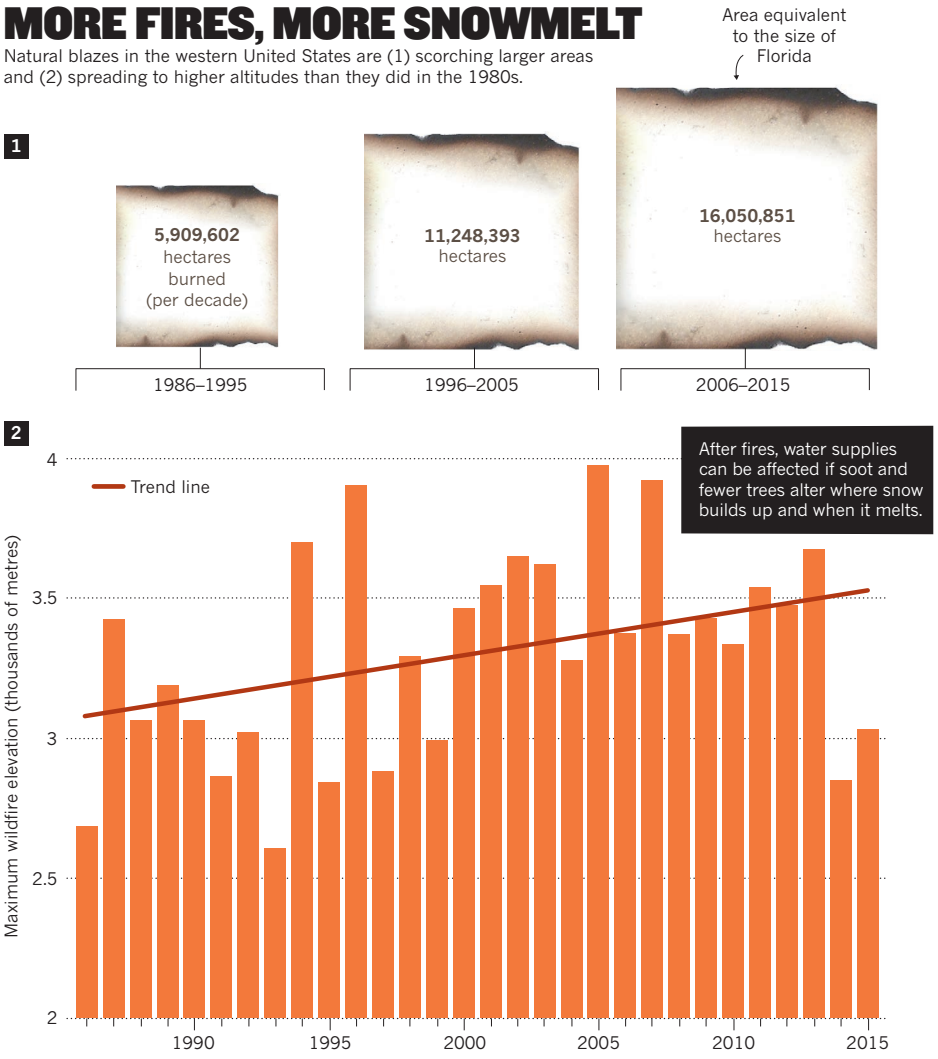
Communities are part of these cycles. For example, 60% of southern California's water comes from melt water from the Sierra Nevada mountains[7]. California's multibillion-dollar agricultural industry also depends on this source. Changing patterns of temperature, snow, wildfires and floods are challenging the state's ageing network of dams, levees and reservoirs. These need to be able to hold more water earlier in the season, as well as prevent floods and debris flows. Minor events that would not normally cause concern can have profound impacts: unexpected bursts of melt water might trigger debris flows over burned land, for instance. Regions in the Andes, Himalayas, Alps and Rockies face similar challenges.

The rapidly changing nature of hazards in a warming world will be unfamiliar to local communities. In Zimbabwe, for example, indigenous peoples are less able to draw on weather patterns, flora and fauna to predict when floods might come and force them to relocate.

Meanwhile, rapid population growth and urbanization exacerbates climate shifts. For example, houses built on steep slopes might become more susceptible to landslides.

MORE FIRES, MORE SNOWMELT

Natural blazes in the western United States are (1) scorching larger areas and (2) spreading to higher altitudes than they did in the 1980s.



SOURCE: MTBS.GOV

MISSING LINKS

Climate researchers have begun to assess some coupled risks, such as from droughts and heatwaves[1,2]. And efforts to minimize human and financial losses from disasters have become more cross-disciplinary and coordinated. The 2015 United Nations Sendai Framework for Disaster Risk Reduction is supporting studies of risk, exposure and vulnerability to improve resilience and emergency responses to a range of disasters, from floods to earthquakes. And the European Union's Seventh Framework Programme (FP7) is improving knowledge, forecasting and decision tools for disaster prevention and intervention.

"A universal framework for addressing cascading disasters is missing."

But these programmes have yet to weave the whole tapestry of hazards together. They typically address one-off major disasters instead of connected chains of smaller events, with a focus on responding to crises rather than preventing them or boosting resilience. Some obvious links can be studied, such as the rapid sequence of events that followed the 2011 Tohoku earthquake, which triggered a tsunami that caused the meltdown of the Fukushima Daiichi nuclear reactor in Japan. But longer-term impacts of the tsunami on the region's hydrology or coasts have not been explored. In addition, most countries adopt their own approaches for dealing with hazards. A universal framework for addressing cascading disasters is missing from current practices.

Many research gaps remain to be filled. Physical cascading mechanisms, such as the impacts of wildfire soot on snowpacks or ocean waves on coastal landslides, and their feedbacks, are poorly understood. Assuming events are independent also gives a false sense of how often these events should be expected to occur, which, in turn, affects disaster preparedness.

Theoretical risk models need to be able to handle multidimensional and interdependent hazards. For example, coastal erosion might be affected by global sea-level rise, earthquake-driven tsunamis, storms

© 2018 Springer Nature Limited. All rights reserved.



JES AZNAR/NYT/REDUX/EYEVINE

A search for survivors after Typhoon Mangkhut triggered a landslide last week in Itogon, the Philippines.

and infrastructure such as barriers and protections. The state of the coasts, in turn, dictates communities' exposure.

Historical records might not hold all the answers. It is also hard to disentangle causes and effects in complex networks, especially in managed systems. Quantifying how an initial event increases or reduces risks from succeeding events, and predicting their timing, is difficult[8]. A levee failure, for example, can have many causes over an extended period, including weakening during drought, extreme rainfall, poor design and inadequate maintenance.

Data are sparse, especially from remote regions such as rugged mountains. Agencies and countries do not always share data. Different disciplines and regions use different definitions. Key observations are missing. For example, the extent of a wildfire and its immediate impacts (deaths, lost houses) are recorded, but damage to soil structure is not commonly noted.

On-the-ground corroboration of satellite data is rare. Many countries restrict the use of their environmental and climate data. Long-term data sets are hard to find, especially in the Middle East, Africa, South America and southeast Asia. Disasters triggered by moderate conditions often go unrecorded. And access to computer facilities and training for processing data is inadequate in many developing countries.

There is no standard protocol for collecting environmental, climate and disaster-impact data. Countries have their own ways of tracking monetary impacts, loss of life and livelihoods. Researchers use different methods for describing droughts, storms, heatwaves and wildfires. For example, a meteorologist might define droughts on the basis of a deficit in precipitation, whereas a hydrologist might describe them on the basis of a change in river run-off.

NEXT STEPS

A global system needs to be developed to assess cascading hazards. Researchers should go beyond calculating the statistics of extreme droughts, floods and wildfires in isolation and delve into their interactions with natural and built environments. International organizations such as the World Climate Research Programme and World Meteorological Organization should take the lead in coordinating research. And intergovernmental agencies, including the European Commission, the US Federal Emergency Management Agency and the United Nations Office for Disaster Risk Reduction, as well as other hazard-focused organizations, should develop a global hazard early-warning system. Regulations and educational materials should also be developed to help engineers, decision-makers and the public to minimize their exposure to compound risks and cascading disasters.

"Outreach and public education are crucial to raise awareness of the potential risks of cascading hazards."

The range of data collected for hazard analysis needs to be expanded, in space and on the ground. Observations should be consistent around the globe and shared openly. We advocate for real-time monitoring, to capture extreme and moderate events as they occur, rather than retrospectively. Such a monitoring network could piggyback on other global environmental tracking efforts, such as the Global Earth Observation System of Systems, the Global Earthquake Model project and NASA's Data Portal. Socio-economic information also needs to be collected about people, livestock, buildings and infrastructure that are at risk.

Data protocols need to be broadened and standardized. And agencies need ways to evaluate progress. They can build on previous efforts to develop metrics for individual events. For example, at a meeting in Nebraska in 2009, the United Nations Convention to Combat Desertification, the US Department of Agriculture, the US National Oceanic and Atmospheric Administration and the US National Drought Mitigation Center convened experts from more than 20 nations to agree on a global set of indices for measuring and forecasting drought — the Lincoln Declaration on Drought Indices.

Engineers, planners and decision-makers need to identify vulnerable infrastructure and ecosystems for monitoring. Local and national governmental agencies and researchers should improve regulations, emergency management, and building codes. After the fires in Greece this year, citizens argued that unregulated construction in woodlands and the absence of an official evacuation plan contributed to the high death toll. Community outreach and public education are crucial to raise awareness of the potential risks of cascading hazards, and to save lives and livelihoods as climate impacts mount. ■

Amir AghaKouchak *is associate professor of civil and environmental engineering,* **Laurie S. Huning**, **Omid Mazdiyasni** *and* **Iman Mallakpour** *are postdoctoral scholars, and* **Felicia Chiang** *is a PhD student at the University of California, Irvine, California, USA.* **Mojtaba Sadegh** *is assistant professor in civil engineering at Boise State University, Boise, Idaho, USA.* **Farshid Vahedifard** *is associate professor of civil and environmental engineering at Mississippi State University, Mississippi, USA.* **Hamed Moftakhari** *is assistant professor of civil, construction and environmental engineering at the University of Alabama, Tuscaloosa, Alabama, USA.*
e-mail: amir.a@uci.edu

1. Chiang, F., Mazdiyasni, O. & AghaKouchak, A. *Sci. Adv.* **4**, eaat2380 (2018).
2. Mazdiyasni, O. & AghaKouchak, A. *Proc. Natl Acad. Sci. USA* **112**, 11484–11489 (2015).
3 Jolly, W. M. *et al. Nature Commun.* **6**, 7537 (2015).
4. Westerling, A. L., Hidalgo, H. G., Cayan, D. R. & Swetnam, T. W. *Science* **313**, 940–943 (2006).
5. Flanner, M. G., Shell, K. M., Barlage, M., Perovich, D. K. & Tschudi, M. A. *Nature Geosci.* **4**, 151–155 (2011).
6. Painter, T. H., Skiles, S. M., Deems, J. S., Brandt, W. T. & Dozier, J. *Geophys. Res. Lett.* **45**, 797–808 (2017).
7. Waliser, D. *et al. Clim. Change* **109** (Suppl. 1), 95–117 (2011).
8. Pescaroli, G. & Alexander, D. *Natural Hazards* **82**, 175–192 (2016).

© 2018 Springer Nature Limited. All rights reserved.



BSIP/UIG VIA GETTY

What should babies be screened for, and who should share in the data?

SOCIETY

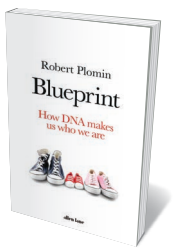
Genetic determinism redux

Nathaniel Comfort questions a psychologist's troubling claims about genes and behaviour.

It's never a good time for another bout of genetic determinism, but it's hard to imagine a worse one than this. Social inequality gapes, exacerbated by climate change, driving hostility towards immigrants and flares of militant racism. At such a juncture, yet another expression of the discredited, simplistic idea that genes alone control human nature seems particularly insidious.

And yet, here we are again with *Blueprint*, by educational psychologist Robert Plomin. Although Plomin frequently uses more civil, progressive language than did his predecessors, the book's message is vintage genetic determinism: "DNA isn't all that matters but it matters more than everything else put together". "Nice parents have nice children because they are all nice genetically." And it's not just any nucleic acid that matters; it is human chromosomal DNA. Sorry, microbiologists, epigeneticists, RNA experts, developmental biologists: you're not part of Plomin's picture.

Crude hereditarianism often re-emerges after major advances in biological knowledge: Darwinism begat eugenics; Mendelism begat worse eugenics. The flowering of medical genetics in the 1950s led to the notorious, now-debunked idea that men with an extra Y chromosome (XYY genotype) were prone to violence. Hereditarian books such as Charles Murray and Richard Herrnstein's *The Bell Curve* (1994) and Nicholas Wade's 2014 *A Troublesome Inheritance* (see N. Comfort *Nature* **513**, 306–307; 2014) exploited their respective scientific and cultural moments, leveraging the cultural authority of science to advance a discredited, undemocratic agenda. Although *Blueprint* is cut from different ideological cloth, the consequences could be just as grave.



Blueprint: How DNA Makes Us Who We Are
ROBERT PLOMIN
Allen Lane (2018)

The scientific advance this time is the genome-wide association study (GWAS). Invented in 1996, GWAS has gained massively in predictive power with the advent of 'polygenic scores', a statistical tool that in recent years has lured social scientists to the genome, with the promise of genetic explanations for complex traits, such as voting behaviour or investment strategies. As Plomin notes, it was something they had been trying to do for a long time.

Plomin's predecessors tried to get monogenic risk scores. For example, Henry Goddard, an educational psychologist who from 1906 to 1918 directed the New Jersey Training School for Feeble-Minded Girls and Boys in Vineland, claimed he had found the gene for low intelligence. With Charles Davenport, a prominent US eugenicist, whispering in his ear, Goddard suggested that learning disabilities resulted from a single Mendelian recessive gene. Scanning the swathes of pedigrees he had collected (progressive-era 'big data'; see *Nature* **558**, 28–29; 2018), he identified what seemed to be a unit character: an apparent recessive "gene for" learning disability. When he factored in behaviours thought to result from that condition — such as criminality and promiscuity — the alleged association went sky-high. Goddard's pedigrees bloomed with antisocial traits, which he believed were passed down ▶

© 2018 Springer Nature Limited. All rights reserved.



EDWIN REMSBERG/VWPICS/REDUX/EYEVINE

Free, healthy school meals have been shown to improve educational attainment

the generations as a Mendelian recessive gene. He never seems to have questioned whether a single gene for such a complex trait made sense biologically. It doesn't.

No one is so foolish as to believe in a single gene for learning disability any more. As has been well established, the genetic contribution to complex traits is spread over many genes, each contributing a minuscule sliver of the variability for the trait. Polygenic risk scores sum and weight these many tiny effects, creating what some researchers have called a "monogenic equivalent"— a "gene for" by proxy.

A polygenic score is a correlation coefficient. A GWAS identifies single nucleotide polymorphisms (SNPs) in the DNA that correlate with the trait of interest. The SNPs are markers only. Although they might, in some cases, suggest genomic neighbourhoods in which to search for genes that directly affect the trait, the polygenic score itself is in no sense causal. Plomin understands this and says so repeatedly in the book — yet contradicts himself several times by arguing that the scores are, in fact, causal.

Plomin deploys a standard feint in hereditarian psychology, insisting on the trivial so-called first law of behavioural genetics: that no psychological trait is entirely unaffected by genetics. But he insists that "genetics is the main systematic force in life", often mediating both gene–environment effects and even environmental effects, such as breastfeeding and TV-watching on school achievement. If all you have is a polygenic score, everything looks like a gene. *Blueprint* is uncritical DNA boosterism, and Plomin "unabashedly a cheerleader" by his own admission.

Polygenic scores do suggest some things to cheer about. We should applaud the broad-based shift across biomedicine from monogenic to polygenic causation. This approach analyses behaviour in a much more complex, surgical way than the crude stabs of Goddard's ilk. The method is finding wide application, from precision medicine to field biology. For example, polygenic scores have been shown to improve risk predictions for prostate, ovarian and breast cancers. They can point to traits that might have been influenced by local adaptation, and gauge the pace of evolutionary change.

"Plomin's argument is just old hereditarian wine pipetted into thousands of tiny polygenic bottles."

Plomin adopts the language of personalized medicine to call for DNA-driven advances in education policy — "personalized learning". He argues that we should think of personality traits as we do autism or attention-deficit hyperactivity disorder: as existing on spectra. He urges psychologists to move away from the language of disorders and to talk instead of graded "dimensions"— personality traits, such as introversion or agreeableness. "All children have special needs," he once told the newspaper *The Guardian*. In a book so filled with retrograde ideas about genes, I was pleasantly surprised to find this strong, welcome biological support for the idea of neurodiversity.

In fundamental ways, however, Plomin's argument is just old hereditarian wine pipetted into thousands of tiny polygenic bottles. In 1969, educational psychologist Arthur Jensen dropped a pseudo-statistical bombshell in the *Harvard Educational Review*. He argued that genetics was responsible for the notional IQ gap between African Americans and white people (not bias baked into the test or environmental effects) and that remedial education was pointless. Jensen's arguments and much of his 'data' were old, part of a dark tradition of hereditarian social science that would subsequently emerge in books such as *The Bell Curve*. *Blueprint* uses language, imagery, rhetoric, conclusions and numbers that will be familiar to readers who have, like me, slogged through all these works. A sobering theme of most, *Blueprint* included, is their aspiration of shaping social policy.

Like much of that literature, *Blueprint* plays fast and loose with the concept of heritability. Sometimes Plomin treats it (correctly) as a variable property of a population in a given environment. As population geneticist

© 2018 Springer Nature Limited. All rights reserved.

Richard Lewontin pointed out in a scathing critique of Jensen's approach in 1970, in times of plenty, height is highly heritable; in a famine, much less so (R. C. Lewontin *Bull. Atom. Sci.* **26**, 2–8; 1970). But elsewhere, Plomin, like Jensen, treats heritability wrongly as a property inherent in a trait.

Blueprint does depart from much prior hereditarian social science in not explicitly mentioning race — the hot-button issue of many earlier works. It instead looks at class. Plomin uses a data set of mostly white British twins, most of whom attended English grammar schools. Yet, given Plomin's extensive experience and his footnotes, the absence of any explicit mention of race (to disavow it, say, or to allude to intersectionality) is conspicuous.

The most troubling thing about *Blueprint* is its Panglossian DNA determinism. Plomin foresees private, direct-to-consumer companies selling sets of polygenic scores to academic programmes or workplaces. Yet, as this "incorrigible optimist" assures us, "success and failure — and credit and blame — in overcoming problems should be calibrated relative to genetic strengths and weaknesses", not environmental ones. All is for the best in this best of brave new worlds.

Plomin likes to say that various components of nurture "matter, but they don't make a difference". But the benefits of good teaching, of school lunches and breakfasts, of having textbooks and air-conditioning and heating and plumbing have been established irrefutably. And they actually are causal: we know why stable blood sugar improves mental concentration. Yet Plomin dismisses such effects as "unsystematic and unstable, so there's not much we can do about them".

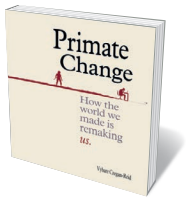
Ultimately, if unintentionally, *Blueprint* is a road map for regressive social policy. Nothing here seems overtly hostile, to schoolchildren or anyone else. But Plomin's argument provides live ammunition for those who would abandon proven methods of improving academic achievement among socio-economically deprived children. His utopia is a forensic world, dictated by polygenic algorithms and the whims of those who know how to use them. People would be defined at birth by their DNA. Expectations would be set, and opportunities, resources and experiences would be doled out — and withheld — a priori, before anyone has had a chance to show their mettle.

To paraphrase Lewontin in his 1970 critique of Jensen's argument, Plomin has made it pretty clear what kind of world he wants.

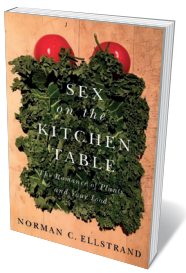
I oppose him. ■

Nathaniel Comfort *is professor of the history of medicine at Johns Hopkins University in Baltimore, Maryland. His most recent book is* The Science of Human Perfection (2012). *He is working on a biography of DNA.*
e-mail: ncomfor1@jhmi.edu

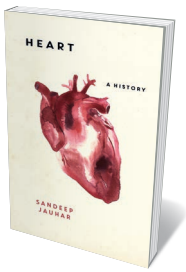
Books in brief



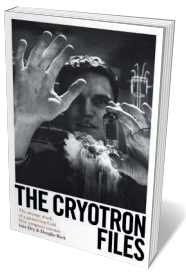
Primate Change: How the World We Made Is Remaking Us
Vybarr Cregan-Reid OCTOPUS *(2018)*
Nature and nurture commingle to fascinating effect in this study of how the environment humans have so thoroughly altered is altering us physiologically. Humanities scholar Vybarr Cregan-Reid ventures from the African forest apes of 20 million years ago to the rise of *Homo sapiens* and the impacts of successive revolutions — agricultural, industrial, urban and digital — on our anatomy. Our grossly sedentary, technologically dominated, polluted present, he argues, constitutes a collective assault on bodies unevolved to cope, leading to 'mismatch' conditions such as myopia and obesity.



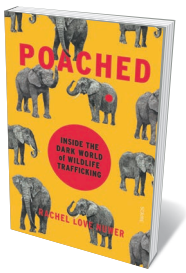
Sex on the Kitchen Table
Norman C. Ellstrand UNIVERSITY OF CHICAGO PRESS *(2018)*
The sex life of an avocado might seem anything but lurid. Geneticist Norman Ellstrand, however, reveals it as a riot of romantic yearning and 'sex switching'. In his foray into the nexus of food, science and plant reproduction, we enter that alternative universe in which olives and quinces are really vehicles for seeds, the tomato (the 'love apple' of yore) is self-fertile and cultivated bananas are female-sterile. You'll become reacquainted with the pistil, and wonder at the sugar beet's rise "from a cascade of geopolitical incidents". Nutrition might never seem the same again.



Heart: A History
Sandeep Jauhar ONEWORLD *(2018)*
Cardiologist Sandeep Jauhar's exploration of that marvellous muscle, the heart, meshes cutting-edge science, memoir and history. He pictures a cadaver's heart as "a squat volcano tipped on its side". He extols physician William Harvey's great 1628 treatise *On the Motion of the Heart and Blood in Animals*. He records the troubled dawn of open-heart surgery, pioneered by experimentalists such as C. Walton Lillehei in the 1950s. And he recounts with raw immediacy his mother's death from cardiac arrest. A moving narrative echoing to the beat of "this organ, prime mover and citadel".



The Cryotron Files
Iain Dey and Douglas Buck ICON *(2018)*
This extraordinary chapter in the annals of cold-war science is both thrilleresque and tragic. At its centre is Dudley Buck, a gifted electrical engineer and US government agent whose prototype microchip, the Cryotron, was key to a covert scheme to create the first supercomputers. As journalist Iain Dey and Buck's son Douglas reveal, Buck and his colleague Louis Ridenour, a physicist, died suddenly in 1959, after a visit from high-level Soviet researchers. Any discussion of Soviet contact-poison hits is speculative; what is not is Buck's substantial contribution to modern computer science.



Poached
Rachel Love Nuwer DACAPO *(2018)*
From the hacked corpses of bull elephants in Botswana to fast-declining pangolin populations, wildlife trafficking is an ongoing threat to conservation gains. Rachel Nuwer, a conservation biologist turned science journalist, traces at first hand the front lines across the globe in her hard-hitting, wince-inducing report. Examining the forces driving demand, the trade itself and countermeasures, she takes us from Africa's killing fields to the corridors of regulatory behemoths, and finds gleams of hope in Chad's National Elephant Action Plan and pangolin rescue efforts in Vietnam. **Barbara Kiser**

© 2018 Springer Nature Limited. All rights reserved.

Correspondence

Mega mine threatens Amazon's rainforests

Citizens of French Guiana in the Amazon rainforest are debating whether the territory should host a huge gold-mining project. A decision from Paris is expected imminently. France's President Emmanuel Macron has declared that such projects will be allowed only if they are exemplary in terms of their environmental and economic ramifications. In our view, this project fails on both counts as currently proposed.

The Russian–Canadian Montagne d'Or mining company plans to extract about 20 kilograms of gold while discarding 80,000 tonnes of matrix every day. Situated between two protected biological reserves, the proposed 800-hectare mining site hosts some 1,558 plant and 546 vertebrate species. The area is also rich in archaeological sites, including 15 sanctuaries built by ancient American Indians.

According to the World Wildlife Fund, the consortium would earn more than €3 billion (US$3.5 billion) over 12 years, of which only 2% would go to French Guiana. The taxes generated by the project would be outpaced by public subsidies, in exchange for just 750 local jobs.

At a time when France is about to officially recognize the importance of preserving biodiversity in its Constitution, it would indeed be paradoxical to authorize this mega-mining project.

François M. Catzeflis*
Montpellier, France.
**On behalf of 5 co-signatories (see go.nature.com/2pnpbnu for complete list).*
francois.catzeflis@wanadoo.fr

Apply data science to benefit society

The translational aspects of data science — the analysis of big data — promise to benefit individuals, science and society. They stand to open up new lines of enquiry in computer science, statistics, ethics, data governance, cognitive psychology, organizational behaviour, information science, sociology and behavioural economics. With an overflowing treasure chest of big data, the time is ripe to tackle the crucial questions that can help translational data science to realize its potential (see, for example, go.nature.com/2nz2qzw and go.nature.com/2kjxa67).

Because it bridges the gap between foundational methods and practical application, translational data science stands to further the study of data-science methods (see D. Donoho *J. Comp. Graph. Stats* **26**, 745–766; 2017). It should also democratize the data-science process and provide knowledge that can inform practical discourse among stakeholders (see also M. Zook *et al. PLoS Comput. Biol.* **13**, e1005399; 2017).

Chaitanya Baru* *University of California, San Diego, California, USA.*
**On behalf of 10 co-signatories (see go.nature.com/2dfgor6 for complete list).*
cbaru@ucsd.edu

Authorship: update to contributorship

The identification of thousands of authors who publish more than 72 papers a year challenges the concept of authorship (J. P. A. Ioannidis *et al. Nature* **561**, 167–169; 2018). Although two of us (G.D.S. and M.K.) fall into this hyperprolific category, we agree.

We suggest that moving from an authorship to a contributorship model (M. Munafò and G. Davey Smith *Nature* **553**, 399–401; 2018) would better reflect the many and varied contributions to large, complex, long-term and management-intensive projects in modern science. Like the credits that roll at the end of a film, the role of each of the many people who contributed would then be recognized (M. K. McNutt *et al. Proc. Natl Acad. Sci. USA* http://doi.org/gc8dnb; 2018). Authors can act as equivalents of film director, script supervisor, second assistant camera, casting director, lead, extra, production accountant or gaffer. As contributors to a research manuscript, they might be credited as, say, data generators, hypothesis constructors, analysts, literature reviewers or evidence-synthesizers.

Today's large teams of physicists and biomedical consortia famously produce high-quality results. The collaborative and multi-disciplinary nature of these huge groups means that their authorship lists are very long. In our view, outdated authorship conventions for such team efforts should be consigned to the past: team science and contributorship are the future.

George Davey Smith, **Marcus Munafò** *MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK.*
Mika Kivimäki *University College London, UK.*
kz.davey-smith@bristol.ac.uk

Authorship: count best, not most

Jobs, promotions and grants are awarded largely on the basis of research productivity. As long as that productivity is defined by quantity (see, for example, J. P. A. Ioannidis *et al. Nature* **561**, 167–169; 2018) rather than quality, the scientific literature is at heightened risk of contamination by inconsequential or even fake results.

There is another way. When scientists are nominated to the US National Academy of Science, or for prizes such as those awarded by the Lasker or Shaw foundations, they are typically asked to summarize their achievements in fewer than 1,000 words and to append a list of up to just 10 supporting publications. In my view, assessments of researchers' performance that similarly focus on their most relevant work and its importance would be more meaningful than simply counting a profusion of thinly sliced papers.

Michele Carbone *University of Hawaii, Honolulu, Hawaii, USA.*
mcarbone@cc.hawaii.edu

Boring speakers talk for longer

Dull talks at conferences can feel interminable. Or could it be that they really do go on for longer?

I investigated this idea at a meeting where speakers were given 12-minute slots. I sat in on 50 talks for which I recorded the start and end time. I decided whether the talk was boring after 4 minutes, long before it became apparent whether the speaker would run overtime. The 34 interesting talks lasted, on average, a punctual 11 minutes and 42 seconds. The 16 boring ones dragged on for 13 minutes and 12 seconds (thereby wasting a statistically significant 1.5 min; t-test, $t = 2.91$, $P = 0.007$). For every 70 seconds that a speaker droned on, the odds that their talk had been boring doubled. For the audience, this is exciting news. Boring talks that seem interminable actually do go on for longer.

To avoid banality, speakers should introduce their objectives early on and focus on pertinent information. They should avoid trite explanations, repetition, getting bogged down by irrelevant minutiae and passing off common knowledge as fresh insight.

Robert M. Ewers *Imperial College London, UK.*
r.ewers@imperial.ac.uk

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at **http://go.nature.com/cmchno**.

© 2018 Springer Nature Limited. All rights reserved.

The role of Olfr78 in the breathing circuit of mice

ARISING FROM A. J. Chang, F. E. Ortega, J. Riegler, D. V. Madison & M. A. Krasnow *Nature* **527**, 240–244 (2015); https://doi.org/10.1038/nature15721

The carotid body is essential for the adaptation of mammals to environmental or pathological conditions that result in hypoxaemia. In response to hypoxia, neuron-like oxygen-sensitive glomus cells in the carotid body release neurotransmitters that rapidly activate afferent sensory fibres that stimulate the respiratory centre and induce hyperventilation[1], although the mechanisms by which glomus cells detect changes in blood oxygen tension remain unclear[2,3]. Single dissociated glomus cells can respond robustly to hypoxia when superfused with standard, lactate-free hypoxic solutions; it was therefore surprising that Chang et al.[4] claimed that lactate activation of the odorant receptor Olfr78, which is abundantly expressed in the carotid body, is required for oxygen regulation of breathing. We are unable to replicate these findings and show that *Olfr78*$^{-/-}$ mice have a normal ventilatory response to hypoxia and that the physiological responses of single glomus cells to hypoxia and lactate are indistinguishable between wild-type and *Olfr78*$^{-/-}$ mice. There is a Reply to this Comment by Chang, A. J. et al. *Nature* **561**, https://doi.org/10.1038/s41586-018-0547-7 (2018).

In addition to being expressed in the main olfactory epithelium, the odorant receptor Olfr78 (also known as MOL2.3 and MOR18-2) is ectopically expressed in the carotid body[4,5] and other mouse tissues, such as medulla oblongata[6], various sympathetic and parasympathetic ganglia[7], and the kidney[8] and colon[9]. This odorant receptor confers responsiveness to short-chain fatty acids[8,9]. Three independent sets of plethysmographic experiments[3,10,11] were performed in Cambridge (UK) with mice carrying an *Olfr78* null allele[12] that were bred and maintained in Frankfurt (Germany), and are henceforth referred to as 'FRA' mice. These experiments revealed a similar ventilatory response to hypoxia in *Olfr78*$^{-/-}$ mice compared with wild-type *Olfr78*$^{+/+}$ or heterozygous *Olfr78*$^{+/-}$ mice (Extended Data Fig. 1a, b). The first set of experiments was carried out before the publication of the previous study[4], and was followed by two sets of experiments performed by the same investigator without knowledge of the genotype. These preliminary findings led us to design an independent plethysmographic study with wild-type and *Olfr78*$^{-/-}$ FRA mice in Seville (Spain), where these whole-animal experiments were followed by analyses of single glomus cell responsiveness to hypoxia in in vitro preparations[1,3]. Wild-type and *Olfr78*$^{-/-}$ littermates were bred and genotyped in Frankfurt, and shipped to Seville without providing information about the genotype; the coat colour was not indicative of the genotype. The genotypes were confirmed in Seville by PCR of genomic DNA (Extended Data Fig. 1c, d) after completion of the plethysmographic and cellular (amperometric) analyses. A second group of FRA mice, of which the genotype was disclosed in advance, was used to study the effect of hypoxia or lactate on cytosolic Ca^{2+} levels in single dissociated glomus cells from wild-type and *Olfr78*$^{-/-}$ mice.

Exposure of FRA mice to hypoxia (10% O_2) or hypercapnia (5% CO_2) elicited characteristic increases in breathing frequency, which did not significantly differ between *Olfr78*$^{-/-}$ and wild-type mice (Fig. 1a, b and Extended Data Fig. 2a–d). In parallel experiments performed independently at Duke University (United States), normal ventilatory responses to hypoxia were observed in homozygous mice carrying a different *Olfr78*-null allele, which had been generated by Lexicon Genetics (Extended Data Fig. 2e, 'LEX' mice).

It cannot be excluded formally that genetic drift, suppressor mutations or other genetic or epigenetic changes might have occurred in the FRA mouse colony. Therefore, in 2017 a new cryorecovery was ordered and the resulting heterozygous mice were shipped directly to Seville and Duke Universities, without passing through the Frankfurt animal facility. These colonies (referred to as 'JAX' mice) were expanded locally and studied after they reached adulthood. Plethysmographic analyses performed in the two laboratories showed similar ventilatory responses to hypoxia and hypercapnia in wild-type and *Olfr78*$^{-/-}$ JAX mice (Fig. 1c, d and Extended Data Fig. 2f–h). Our results contrast markedly with the previous study[4], which reported a complete abolition of the ventilatory response to hypoxia in *Olfr78*$^{-/-}$ JAX mice. Their mice carried the same *Olfr78*-null allele[12], and had also been obtained from The Jackson Laboratory, without passing through Frankfurt.

To investigate the oxygen-sensing properties of Olfr78-deficient glomus cells, tissue slices or enzymatically dissociated cells were obtained from the carotid bodies that were dissected out from wild-type and

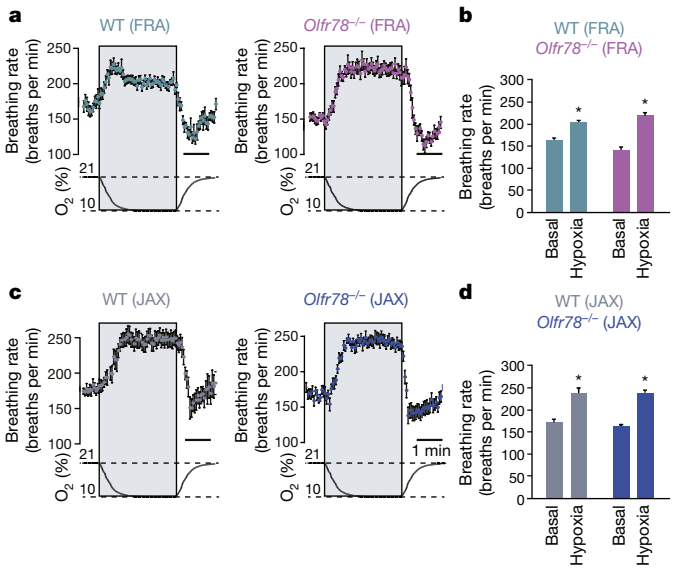


Fig. 1 | Ventilatory responses to hypoxia of wild-type and *Olfr78*$^{-/-}$ mice. a, Plethysmographic recordings (breathing frequency as a measure of time) of the ventilatory response to hypoxia (10% O_2) performed on wild-type (WT; $n = 10$) and *Olfr78*$^{-/-}$ ($n = 10$) FRA mice. Each data point represents the mean ± s.e.m. of the values for the group of 10 mice. Oxygen (percentage O_2) tensions are indicated at the bottom. **b**, Breathing frequency in normoxia (21% O_2) and during exposure to hypoxia (10% O_2) in *Olfr78*$^{-/-}$ FRA mice ($n = 10$) compared to their wild-type littermates ($n = 10$). **c**, **d**, Similar experiments as in **a**, **b** performed with wild-type mice ($n = 10$) and *Olfr78*$^{-/-}$ mice ($n = 10$) from the JAX colony. Data are mean ± s.e.m. Statistically significant differences compared to basal values are indicated; $^*P < 0.001$.

© 2018 Springer Nature Limited. All rights reserved.

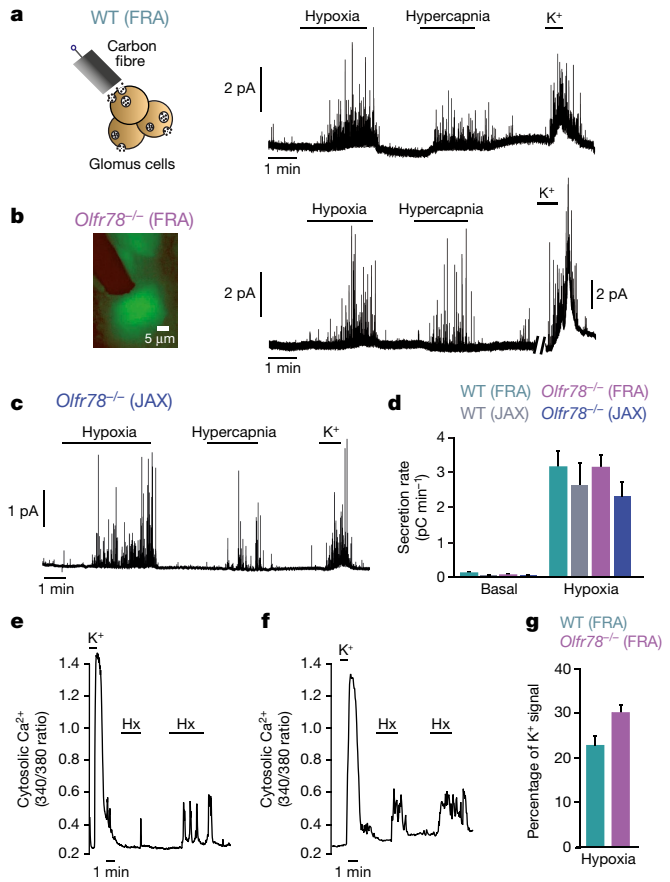


Fig. 2 | Physiological responses to acute hypoxia in carotid body slices and single dissociated glomus cells from wild-type and *Olfr78*$^{-/-}$ mice. **a**–**c**, Amperometric recordings showing glomus cell secretory responses to hypoxia ($p_{O_2} \approx 15$ mm Hg), hypercapnia (5% CO_2) and 40 mM KCl in carotid body slices from wild-type (**a**) and *Olfr78*$^{-/-}$ (**b**) mice of the FRA colony, as well as *Olfr78*$^{-/-}$ JAX mice (**c**). **a**, The drawing on the left illustrates the detection of exocytotic dopamine release by amperometry using a polarized carbon-fibre electrode. **b**, The microphotograph shows the carbon-fibre electrode and the intrinsic GFP fluorescence of glomus cells from *Olfr78*$^{-/-}$ mice. **d**, Secretion rate (picocoulombs per min) in basal conditions (normoxia, $p_{O_2} \approx 145$ mm Hg) and in response to hypoxia ($p_{O_2} \approx 15$ mm Hg) in carotid body slices from wild-type FRA mice (green; $n = 23$ cells, 9 mice) and wild-type JAX mice (grey; $n = 12$ cells, 7 mice) as well as *Olfr78*$^{-/-}$ FRA mice (purple; $n = 19$ cells, 7 mice) and *Olfr78*$^{-/-}$ JAX mice (blue; $n = 11$ cells, 6 mice). **e**, **f**, Increases in cytosolic Ca^{2+} levels elicited by hypoxia (Hx; approximately 15 mm Hg) in single, dissociated, Fura-2-loaded glomus cells from wild-type (**e**) and *Olfr78*$^{-/-}$ (**f**) FRA mice. **g**, Amplitude of changes in cytosolic Ca^{2+} levels induced by hypoxia in glomus cells from wild-type FRA mice (green; $n = 17$ cells, 5 mice) and *Olfr78*$^{-/-}$ FRA mice (purple; $n = 25$ cells, 5 mice) relative to the signal obtained with 40 mM K^+. Data are expressed as mean ± s.e.m.

Olfr78$^{-/-}$ mice. It is well-established that glomus cells, which can be activated by hypoxia, hypercapnia and other stimuli, are responsible for the chemosensory properties of the carotid body[1]. Cells in slices or dissociated glomus cells exhibit increases in exocytotic dopamine release or cytosolic [Ca^{2+}] as an inverse function of oxygen tension (p_{O_2}), which parallel the relationship between ventilation and p_{O_2} observed in the entire animal[1,13,14]. We fixed slices of the carotid body after studying the cellular responses to hypoxia, and determined co-expression of GFP (contained in the gene-targeted *Olfr78* allele[4,10]; Extended Data Fig. 1c) and tyroxine hydroxylase (TH), a characteristic marker of glomus cells (Extended Data Fig. 3a). In the initial in vitro experiments, glomus cells from wild-type and *Olfr78*$^{-/-}$ FRA mice showed similar secretory responses to hypoxia, hypercapnia and high K^+, as measured by amperometry using a polarized carbon fibre electrode, placed near the cells to oxidize the dopamine molecules that were released[3,14] (Fig. 2a, b). A normal responsiveness to hypoxia and the other stimuli was confirmed in glomus cells from *Olfr78*$^{-/-}$ JAX mice (Fig. 2c). The secretion rate (a measure of the number of dopamine molecules released per unit of time)[1,3] during exposure to hypoxia were the same in glomus cells from wild-type and *Olfr78*$^{-/-}$ mice of the FRA and JAX colonies (Fig. 2d). In addition, single dissociated glomus cells of wild-type and *Olfr78*$^{-/-}$ FRA mice, loaded with Fura-2[3,14], showed no difference in their increases of cytosolic Ca^{2+} levels to hypoxia (Fig. 2e–g). Therefore, our results at the cellular level are consistent with our results at the whole-animal level. We confirmed by quantitative PCR the absence of *Olfr78* mRNA in superior cervical ganglion tissue obtained from *Olfr78*$^{-/-}$ mice. The mRNA levels normalized to wild-type mice were: *Olfr78*$^{+/+}$, 1 ± 0.18, $n = 4$; *Olfr78*$^{+/-}$, 0.52 ± 0.10, $n = 5$; *Olfr78*$^{-/-}$, 0.0 ± 0.0, $n = 5$.

We thus cannot confirm the main conclusion reached in the previous study[4]. Genetic drift in a cohort of mice used by Chang et al. could have caused a loss of sensitivity of the carotid body to hypoxia in a way that is unrelated to the *Olfr78* mutation. Another possibility is that environmental factors, diet or co-existing microbiome that is specific to the animal facility in Stanford at the time of the experiments could have modulated the glomus cell function to become dependent on Olfr78. However, we obtained highly consistent data using several sources of mice bred or maintained in animal facilities in four countries (Spain, Germany, the United Kingdom and the United States).

The absolute requirement of lactate for carotid body activation during hypoxia suggested by Chang et al.[4] is incompatible with abundant data demonstrating that acute responsiveness to hypoxia is a cell-autonomous phenomenon, maintained in single dissociated glomus cells perfused with lactate-free solutions[1,3,13,14] (Fig. 2e–g). Unfortunately, Chang et al. did not support their conclusion with recordings of single cells. We have also observed that lactate at relatively high concentrations (around 5 mM or higher) can activate glomus cells from wild-type as well as *Olfr78*$^{-/-}$ FRA mice (Extended Data Fig. 3b–d). It seems, therefore, that Olfr78 is not required to confer lactate responsiveness to glomus cells. Two recent studies reported that lactate is a poor agonist for Olfr78[5,15], using the same heterologous expression system that was used in Chang et al.[4].

In summary, we conclude that Olfr78 is not required for the ventilatory response to hypoxia in mice, and is not required for activation of carotid body cells by lactate.

Data availability

Data are available upon request from the corresponding author.

Hortensia Torres-Torrelo[1,2], Patricia Ortega-Sáenz[1,2], David Macías[3], Masayo Omura[4], Ting Zhou[5], Hiroaki Matsunami[5,6], Randall S. Johnson[3,7], Peter Mombaerts[4] & José López-Barneo[1,2]*

[1]Instituto de Biomedicina de Sevilla (IBiS), Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville, Spain. [2]Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Seville, Spain. [3]Department of Physiology, Development & Neuroscience, University of Cambridge, Physiological Laboratory, Cambridge, UK. [4]Max Planck Research Unit for Neurogenetics, Frankfurt, Germany. [5]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC, USA. [6]Department of Neurobiology and Duke Institute for Brain Sciences, Duke University Medical Center, Durham, NC, USA. [7]Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden. *e-mail: lbarneo@us.es

© 2018 Springer Nature Limited. All rights reserved.

Received: 13 February 2017; Accepted: 7 June 2018;
Published online: 26 September 2018

1. López-Barneo, J. et al. Oxygen-sensing by arterial chemoreceptors: mechanisms and medical translation. *Mol. Aspects Med.* **47–48**, 90–108 (2016).
2. Peers, C. Acute oxygen sensing—inching ever closer to an elusive mechanism. *Cell Metab.* **22**, 753–754 (2015).
3. Fernández-Agüera, M. C. et al. Oxygen sensing by arterial chemoreceptors depends on mitochondrial complex I signaling. *Cell Metab.* **22**, 825–837 (2015).
4. Chang, A. J., Ortega, F. E., Riegler, J., Madison, D. V. & Krasnow, M. A. Oxygen regulation of breathing through an olfactory receptor activated by lactate. *Nature* **527**, 240–244 (2015).
5. Zhou, T., Chien, M.-S., Kaleem, S. & Matsunami, H. Single cell transcriptome analysis of mouse carotid body glomus cells. *J. Physiol.* **594**, 4225–4251 (2016).
6. Conzelmann, S. et al. A novel brain receptor is expressed in a distinct population of olfactory sensory neurons. *Eur. J. Neurosci.* **12**, 3926–3934 (2000).
7. Weber, M., Pehl, U., Breer, H. & Strotmann, J. Olfactory receptor expressed in ganglia of the autonomic nervous system. *J. Neurosci. Res.* **68**, 176–184 (2002).
8. Pluznick, J. L. et al. Olfactory receptor responding to gut microbiota-derived signals plays a role in renin secretion and blood pressure regulation. *Proc. Natl Acad. Sci. USA* **110**, 4410–4415 (2013).
9. Fleischer, J., Bumbalo, R., Bautze, V., Strotmann, J. & Breer, H. Expression of odorant receptor Olfr78 in enteroendocrine cells of the colon. *Cell Tissue Res.* **361**, 697–710 (2015).
10. Macías, D., Fernández-Agüera, M. C., Bonilla-Henao, V. & López-Barneo, J. Deletion of the von Hippel–Lindau gene causes sympathoadrenal cell death and impairs chemoreceptor-mediated adaptation to hypoxia. *EMBO Mol. Med.* **6**, 1577–1592 (2014).
11. Bleymehl, K. et al. A sensor for low environmental oxygen in the mouse main olfactory epithelium. *Neuron* **92**, 1196–1203 (2016).
12. Bozza, T. et al. Mapping of class I and class II odorant receptors to glomerular domains by two distinct types of olfactory sensory neurons in the mouse. *Neuron* **61**, 220–233 (2009).
13. Buckler, K. J. & Vaughan-Jones, R. D. Effects of hypoxia on membrane potential and intracellular calcium in rat neonatal carotid body type I cells. *J. Physiol.* **476**, 423–428 (1994).
14. Ureña, J., Fernández-Chacón, R., Benot, A. R., Álvarez de Toledo, G. A. & López-Barneo, J. Hypoxia induces voltage-dependent Ca^{2+} entry and quantal dopamine secretion in carotid body glomus cells. *Proc. Natl Acad. Sci. USA* **91**, 10208–10211 (1994).
15. Aisenberg, W. H. et al. Defining an olfactory receptor function in airway smooth muscle cells. *Sci. Rep.* **6**, 38231 (2016).

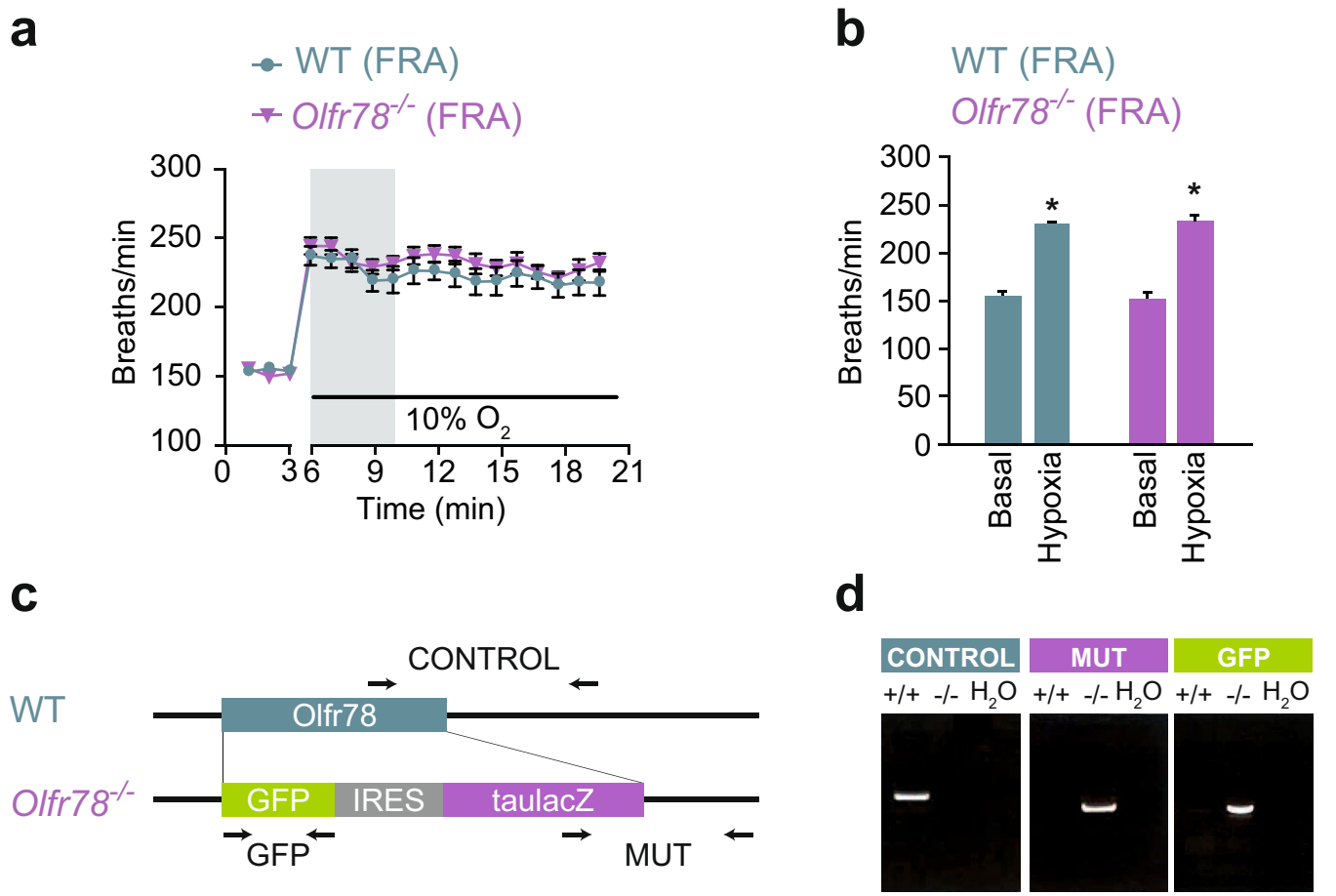
Author contributions All authors participated in the design of the experiments. H.T.-T., P.O.-S., D.M., M.O. and T.Z. performed the experiments and analysed data. R.S.J., J.L.-B., H.M. and P.M. supervised the study and wrote the manuscript.

Competing interests Declared none.

Additional information
Extended data accompanies this Comment.
Supplementary information accompanies this Comment.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.L.-B.

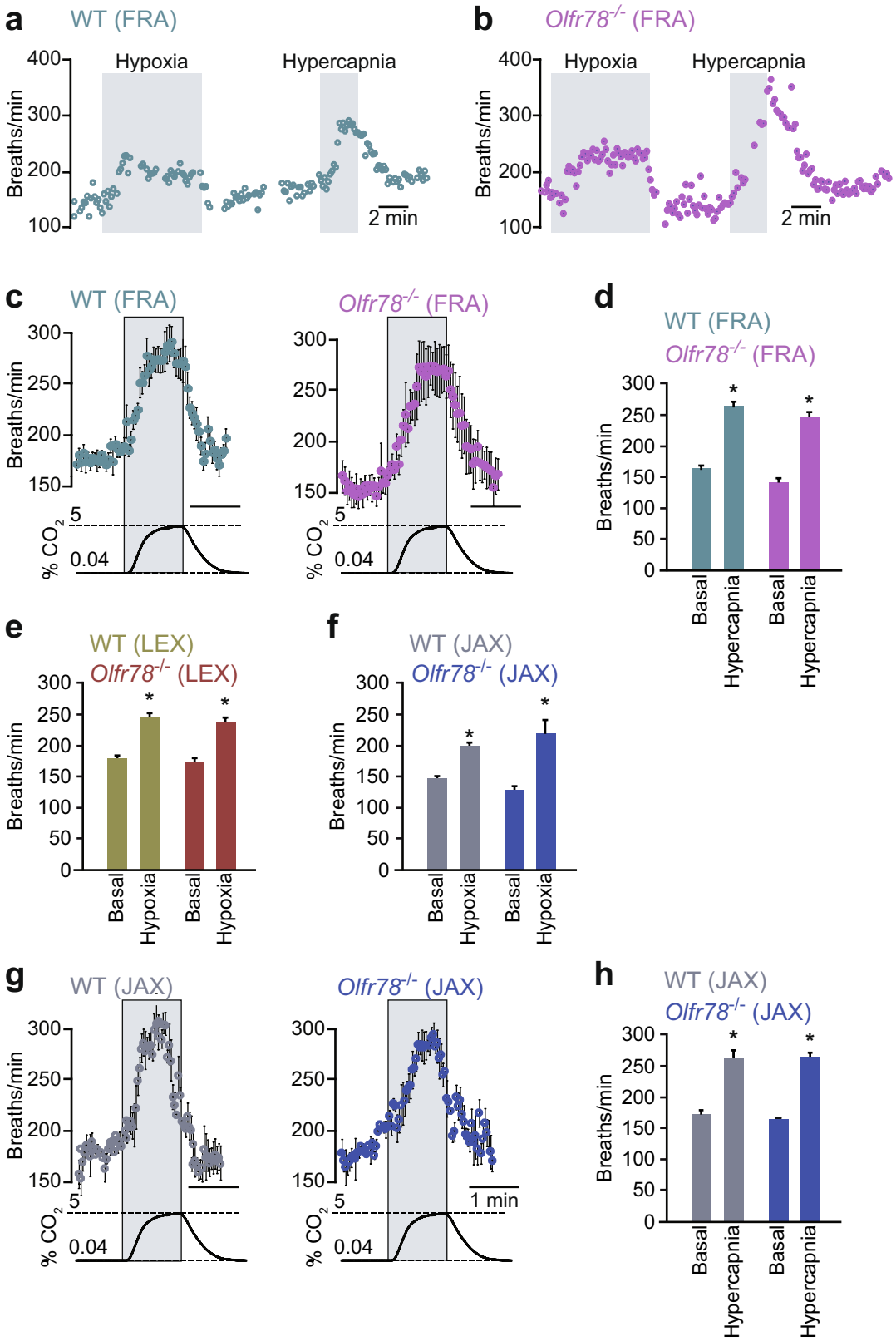
https://doi.org/10.1038/s41586-018-0545-9

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Ventilatory responses to hypoxia and genotyping in wild-type and *Olfr78*$^{-/-}$ mice. a, Plethysmographic recordings (breathing frequency as a measure of time) of the ventilatory response to hypoxia (10% O_2). Each data point represents the mean ± s.e.m. of the values for control mice ($n = 10$; 7 wild-type and 3 heterozygous mice pooled together) and *Olfr78*$^{-/-}$ mice ($n = 9$) of the FRA colony. The grey-shaded area indicates the first five consecutive measurements after the transition to hypoxia used for quantification (see Supplementary Information). **b**, Breathing frequency in normoxia (basal) and during exposure to hypoxia (10% O_2) of control FRA mice ($n = 10$, 7 wild-type and 3 heterozygous mice) and *Olfr78*$^{-/-}$ FRA mice ($n = 9$). Data are mean ± s.e.m. Statistically significant differences compared to basal values are indicated; $^*P < 0.001$. **c**, Schematic of the wild-type and gene-targeted *Olfr78* alleles. Olfr78, intronless coding region of *Olfr78*; GFP, green-fluorescent protein; IRES, internal ribosome entry site; taulacZ, fusion of bovine tau with β-galactosidase. The arrows indicate the position and orientation of PCR primers used for genotyping. **d**, Genotyping of genomic tail DNA of wild-type (*Olfr78*$^{+/+}$) and homozygous (*Olfr78*$^{-/-}$) mice by PCR. The PCR primer pair 'CONTROL' amplifies the wild-type *Olfr78* allele; the PCR primer pair 'GFP' amplifies internal *GFP* sequences; and the PCR primer pair 'MUT' amplifies the gene-targeted *Olfr78* allele.

© 2018 Springer Nature Limited. All rights reserved.

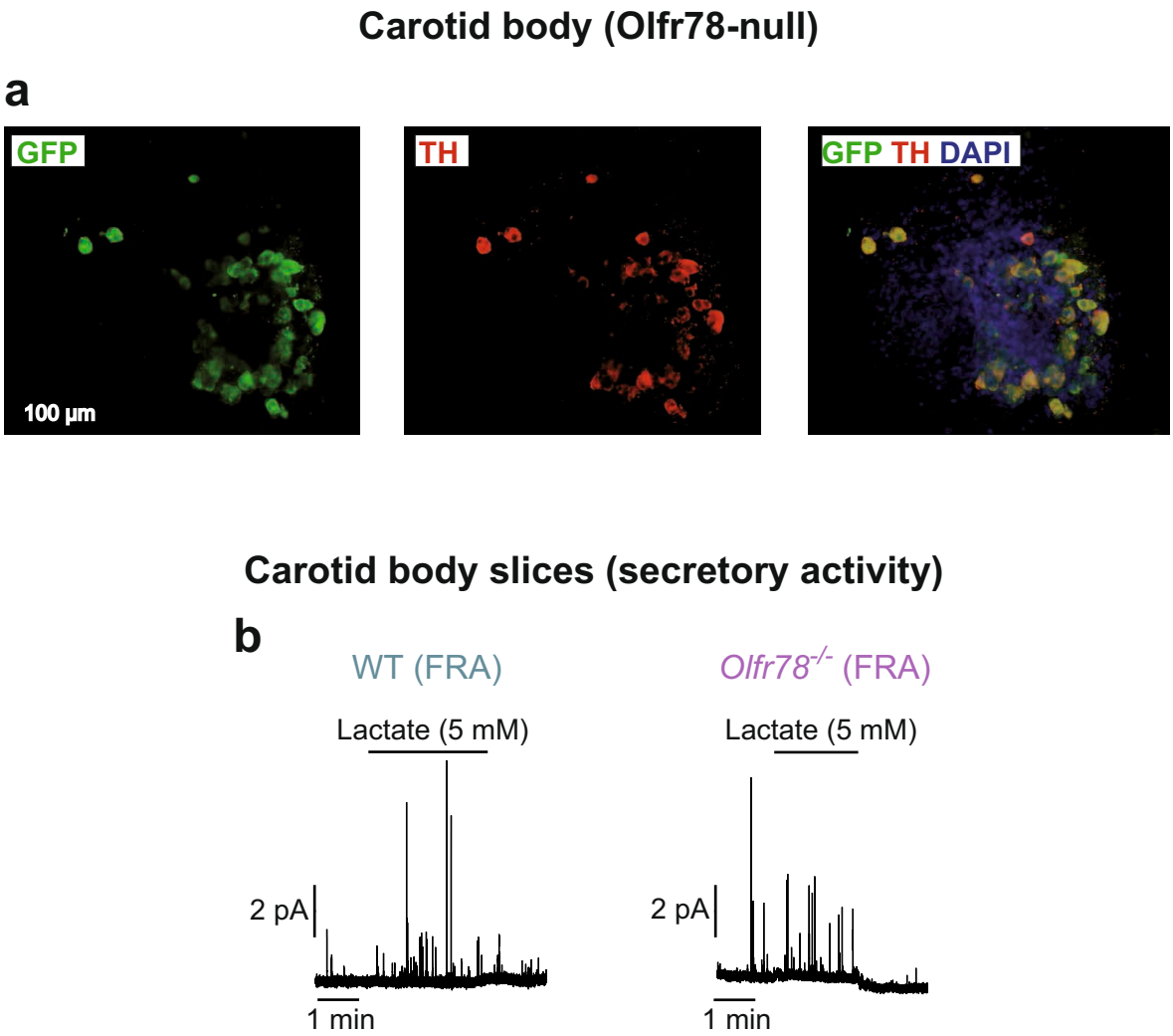


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 2 | Ventilatory responses to hypoxia and hypercapnia in wild-type and *Olfr78*$^{-/-}$ mice. a, **b**, Representative examples of plethysmographic recordings (breathing frequency) during exposure to hypoxia (10% O_2) and hypercapnia (5% CO_2) in a wild-type FRA mouse and in a *Olfr78*$^{-/-}$ FRA mouse. **c**, Plethysmographic recordings (breathing frequency as a measure of time) of the ventilatory response to hypercapnia (5% CO_2) performed on wild-type ($n = 10$) and *Olfr78*$^{-/-}$ ($n = 10$) FRA mice. Each data point represents the mean ± s.e.m. of the values for the group of 10 mice. CO_2 (percentage CO_2) tensions are indicated at the bottom. **d**, Breathing frequency during exposure to hypercapnia (5% CO_2) in *Olfr78*$^{-/-}$ FRA mice ($n = 10$) compared to their wild-type littermates ($n = 10$). **e**, Breathing frequency during exposure to hypoxia (10% O_2) in *Olfr78*$^{-/-}$ LEX mice compared to wild-type LEX mice ($n = 10$ for each genotype, 7 pairs in a C57BL/6 background, 3 pairs in a C57BL/6:129S5 mixed background, 9 out of 10 pairs are sex-matched littermates). **f**, Breathing frequency during exposure to hypoxia (10% O_2) in *Olfr78*$^{-/-}$ JAX mice ($n = 6$) compared to their wild-type littermates ($n = 5$), in experiments carried out at Duke University. **g**, Plethysmographic recordings (breathing frequency as a measure of time) of the ventilatory response to hypercapnia (5% CO_2) performed on wild-type ($n = 7$) and *Olfr78*$^{-/-}$ ($n = 7$) JAX mice. Each data point represents the mean ± s.e.m. CO_2 (percentage CO_2) tensions are indicated at the bottom. **h**, Breathing frequency during exposure to hypercapnia (5% CO_2) in *Olfr78*$^{-/-}$ JAX mice ($n = 7$) compared to their wild-type littermates ($n = 7$). Data are mean ± s.e.m. Statistically significant differences compared to basal values are indicated; $^*P < 0.001$.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 3 | Changes in cellular parameters elicited by lactate in carotid body glomus cells from wild-type and *Olfr78*$^{-/-}$ FRA mice. **a**, Immunohistochemical detection of GFP and tyrosine hydroxylase (TH) in a carotid body slice from an *Olfr78*$^{-/-}$ FRA mouse. **b**, Representative examples of quantal dopamine secretion from glomus cells in carotid body slices of wild-type FRA mice (left) and *Olfr78*$^{-/-}$ FRA mice (right) in response to external application of L-lactate (sodium lactate, 5 mM). **c**, Representative examples of increase in cytosolic Ca^{2+} levels elicited by L-lactate (5 mM) in single dissociated glomus cells from wild-type FRA mice (left) and *Olfr78*$^{-/-}$ FRA mice (right). **d**, Amplitude of changes in cytosolic Ca^{2+} levels induced by lactate in glomus cells from wild-type FRA mice (green, $n = 16$ cells, 5 mice) and *Olfr78*$^{-/-}$ FRA mice (purple, $n = 14$ cells, 5 mice) relative to the signal obtained with 40 mM K^{+}. Data are mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

Chang et al. reply

REPLYING TO H. Torres-Torrelo et al. *Nature* **561**, https://doi.org/10.1038/s41586-018-0545-9 (2018)

In the accompanying Comment, Torres-Torrelo et al.[1], and others previously[2], confirmed the following key findings of our paper[3]: *Olfr78* is highly and selectively expressed in carotid body glomus cells; Olfr78 can be activated by lactate, as measured in *Olfr78*-transfected cultured cells; and lactate activates glomus cells at low-millimolar concentrations, within the physiological concentration range of lactate and the range that we measured for lactate activation (half-maximum effective concentration (EC_{50}) of around 4 mM) of *Olfr78*-transfected cultured cells.

The major difference that can be compared is in the observed behavioural response of $Olfr78^{-/-}$ mutant mice to acute hypoxia (hypoxic ventilatory response). The $Olfr78^{tm1Mom}$ mutants used in our original study were cryorecovered by JAX in 2010 from sperm in a mixed 129P2/OlaHsd × C57BL/6J background[3,4], but a marked decline in breeding in heterozygous intercrossed mice within three years of obtaining the animals prevented us from retesting the same cohort now. To address differences between the results presented by Torres-Torrelo et al.[1] and our previous study[3], we instead tested female progeny of newly cryorecovered $Olfr78^{+/-}$ mutant mice from JAX as well as $Olfr78^{+/-}$ mutant mice from the Frankfurt colony using behavioural protocols similar to our original study and the protocol used by Torres-Torrelo et al.[1]. Surprisingly, we did not observe a consistent defect in the hypoxic ventilatory response of either current strain under these conditions. Although we cannot exclude an effect due to differences in our current versus previous test conditions (the animal facility and equipment used in our original experiments in 2011 are no longer available), it seems more likely that the original cohort had an unusual genetic background that rendered a defect in this response more apparent in $Olfr78^{-/-}$ mutants. Because the carotid body response to hypoxia is still partially intact in $Olfr78^{-/-}$ mutants, including a fully functional response to acid[3] that contributes to carotid body oxygen sensing[5], the behavioural response to hypoxia is likely to be highly sensitive to differences in assay conditions and genetic background[6], as observed for mutants in the acid-sensing pathway[5,7]. We proposed that glomus cells separately sense both the extracellular lactate and hydrogen ions from lactic acid produced during acute hypoxia[3]; it may be necessary to disrupt both pathways simultaneously and acutely to observe a robust effect in behavioural assays.

Although we did not find a consistent defect in the hypoxic ventilatory response in the current *Olfr78* mutant strains, curiously we did observe a greater variance in their ventilatory frequency in hypoxia compared to wild-type littermates. Therefore, in addition to potential functional redundancy between Olfr78 and acid-sensing pathways, there may be a variable compensatory response induced by *Olfr78* loss. This could involve either of these carotid body pathways, or the peripheral and central adaptation mechanisms that restore a hypoxic ventilatory response following bilateral carotid body denervation[8,9].

The results of Torres-Torrelo et al.[1] from carotid body slices and isolated cells cannot be directly compared to our results derived from monitoring the integrated sensory output of the intact carotid body by recordings of the carotid sinus nerve. There is growing evidence that sensory signalling involves communication between glomus cells and other carotid body cells[10], so the effect of Olfr78 on signalling may be apparent only in the intact organ. Furthermore, Torres-Torrelo et al.[1] measured release of dopamine, which is not an excitatory transmitter in the rodent carotid body[11], and calcium responses of the glomus cells. Olfr78 may regulate a signalling step in glomus cells that is downstream or parallel to these responses, so the observations of Torres-Torrelo et al.[1] are not incompatible with our nerve recordings. A similar pair of seemingly contradictory observations was previously described for acid-sensing mutants, in which the carotid sinus nerve response to hypoxia was partially defective in mutants but the dopamine response was intact[5,12]. It will be important to determine not only how Olfr78 activity in glomus cells is modulated by lactate produced under hypoxia, but also which signalling step(s) the activated receptor regulates.

We thank Torres-Torrelo et al.[1] for their experiments showing that the *Olfr78* mutant phenotype is more complex than our original experiments revealed, and for motivating us and others to carefully define the relationship of Olfr78 to other pathways implicated in oxygen sensing in the carotid body and the physiological and genetic compensatory mechanisms that can influence them and other aspects of the hypoxic ventilatory response.

Three of the listed authors (N.S.K., H.H. and A.D.) contributed only to the work contained in this Reply.

Andy J. Chang[1]*, Noah S. Kim[1], Homza Hireed[1], Alex Diaz de Arce[2], Fabian E. Ortega[2], Johannes Riegler[3], Daniel V. Madison[4] & Mark A. Krasnow[2]*

[1]Department of Physiology and Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA, USA. [2]Department of Biochemistry, Stanford University School of Medicine and Howard Hughes Medical Institute, Stanford, CA, USA. [3]Genentech Inc., South San Francisco, CA, USA. [4]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA.
*e-mail: andy.chang@ucsf.edu; krasnow@stanford.edu26 September 2018

1. Torres-Torrelo, H. et al. The role of Olfr78 in the breathing circuit of mice. *Nature* **561**, https://doi.org/10.1038/s41586-018-0545-9 (2018).
2. Zhou, T., Chien, M. S., Kaleem, S. & Matsunami, H. Single cell transcriptome analysis of mouse carotid body glomus cells. *J. Physiol.* **594**, 4225–4251 (2016).
3. Chang, A. J., Ortega, F. E., Riegler, J., Madison, D. V. & Krasnow, M. A. Oxygen regulation of breathing through an olfactory receptor activated by lactate. *Nature* **527**, 240–244 (2015).
4. Bozza, T. et al. Mapping of class I and class II odorant receptors to glomerular domains by two distinct types of olfactory sensory neurons in the mouse. *Neuron* **61**, 220–233 (2009).
5. Trapp, S., Aller, M. I., Wisden, W. & Gourine, A. V. A role for TASK-1 (KCNK3) channels in the chemosensory control of breathing. *J. Neurosci.* **28**, 8844–8850 (2008).
6. Tankersley, C. G., Fitzgerald, R. S. & Kleeberger, S. R. Differential control of ventilation among inbred strains of mice. *Am. J. Physiol.* **267**, R1371–R1377 (1994).
7. Jungbauer, S. et al. Sex-dependent differences in the in vivo respiratory phenotype of the TASK-1 potassium channel knockout mouse. *Respir. Physiol. Neurobiol.* **245**, 13–28 (2017).
8. Gourine, A. V. & Funk, G. D. On the existence of a central respiratory oxygen sensor. *J. Appl. Physiol.* **123**, 1344–1349 (2017).
9. Hodges, M. R. & Forster, H. V. Respiratory neuroplasticity following carotid body denervation: central and peripheral adaptations. *Neural Regen. Res.* **7**, 1073–1079 (2012).
10. Murali, S. & Nurse, C. A. Purinergic signalling mediates bidirectional crosstalk between chemoreceptor type I and glial-like type II cells of the rat carotid body. *J. Physiol.* **594**, 391–406 (2016).
11. Kumar, P. & Prabhakar, N. R. Peripheral chemoreceptors: function and plasticity of the carotid body. *Compr. Physiol.* **2**, 141–219 (2012).
12. Ortega-Sáenz, P. et al. Carotid body chemosensory responses in mice deficient of TASK channels. *J. Gen. Physiol.* **135**, 379–392 (2010).

Author contributions A.J.C. and M.A.K. designed experiments and wrote the manuscript. A.J.C., N.S.K., H.H. and A.D.d.A. performed experiments and analysed data. All authors discussed and reviewed the manuscript.

Competing interests Declared none.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.J.C. or M.A.K.

https://doi.org/10.1038/s41586-018-0547-7

© 2018 Springer Nature Limited. All rights reserved.

MATERIALS SCIENCE

Flexible self-powered biosensors

Current biological sensors require bulky external power sources. Ultrathin solar cells have now been fabricated that can power flexible, wearable sensors for the precise and continuous monitoring of biological signals. SEE LETTER P.516

SHIMING ZHANG & FABIO CICOIRA

Flexible electronic devices are emerging as powerful tools for measuring biological signals, such as a person's heart rate or blood pressure. This is because these devices are lightweight, can bind to human skin and can tolerate mechanical deformation[1,2]. However, although various flexible biosensors have already been demonstrated, most of them rely on large external power sources to operate. On page 516, Park *et al.*[3] report ultraflexible biosensors, used to monitor heart rate, that are powered by ultrathin solar cells. The integration of ultraflexible power sources and sensors has the potential to revolutionize the technology of self-powered conformable biosensors for applications in wearable electronics and diagnostics.

Park and colleagues achieved this feat by combining solar cells, known as organic photovoltaic (OPV) cells[4], with electronic devices called organic electrochemical transistors (OECTs)[5,6]. The authors fabricated the OPV cells and OECTs on an ultrathin substrate, which was made of a type of plastic known as parylene. The OPV cells could convert up to 10.5% of the energy received from light into electricity — which is among the highest reported values of power-conversion efficiency for ultraflexible devices.

Flexible OPV cells are normally less efficient than their rigid counterparts, which use more-established fabrication processes. Park *et al.* overcame this limitation by introducing zinc oxide structures into the OPV cells. These structures consisted of nanometre-scale patterns that facilitated electron transport in the OPV cells, maximizing their efficiency.

Park and co-workers produced the nanopatterns using a blank DVD, which has these patterns on its surface to store information. The authors first replicated the DVD's nanopatterns onto an elastic stamp. They then used this stamp to generate nanopatterns on the OPV cells by a fabrication technique known as soft lithography[7].

Another key advantage of Park and colleagues' OPV cells over rigid solar cells is that the power-conversion efficiency is insensitive to the angle at which the cells are illuminated by light. In conventional solar cells, light that comes in at a greater angle to the surface of the cells undergoes more reflection, leading to a lower efficiency. But in the authors' devices, the nanopatterns minimize the reflection of incoming light, regardless of the illumination angle. As a result, the efficiency of these devices is unaffected by movement, which is a desirable property for wearable biosensors.

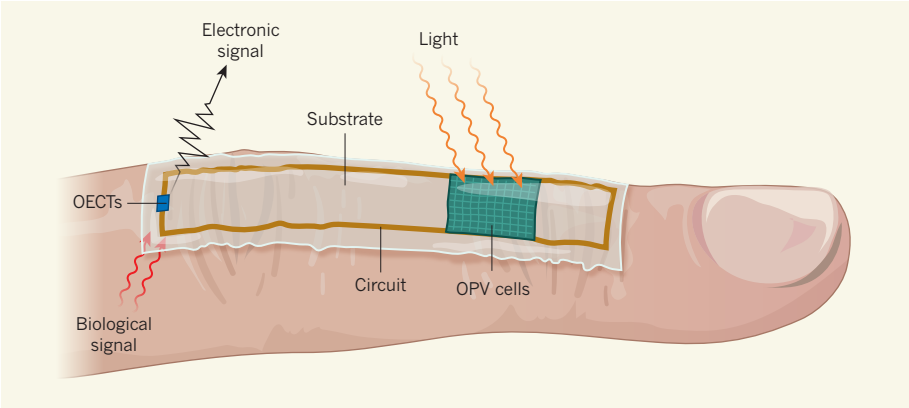


Figure 1 | A self-powered ultraflexible biosensor. Park *et al.*[3] demonstrate a platform for capturing biological signals and converting them into electronic signals, which can then be analysed. The platform is ultraflexible, and does not require external power connections. It consists of electronic devices known as organic electrochemical transistors (OECTs) and solar cells called organic photovoltaic (OPV) cells. These components are connected by an electric circuit and are contained on an ultrathin plastic substrate. The platform is powered by the electrical energy produced when the OPV cells are irradiated with light. The authors attached the platform to a person's finger and a gel electrode to the person's chest (not shown), and found that the platform could be used to monitor the person's heart rate.

The interest in using flexible OPV cells to power flexible sensors naturally calls for stable electrical performance under mechanical deformation. Conventional flexible OPV cells do not meet this requirement because they are comprised of thick, rigid materials, which make the devices fragile. Park *et al.* took advantage of the ultrathin nature of their nanopatterned OPV cells and laminated the devices on a pre-stretched elastomer[8] (a rubber-like material). They found that the resulting devices could not only be laid on curved surfaces, but could also be stretched to twice their initial length (a mechanical strain of 200%) and still maintain high power-conversion efficiency. Even after 900 cycles of stretching and releasing the devices, the efficiency dropped to only about 75% of its initial value.

Park and colleagues used the nanopatterned OPV cells to drive OECTs, which acted as sensitive and flexible biosensors. Such OECTs are able to work using a low voltage (about 1 volt; ref. 9), which is well within the powering capacity of the OPV cells. Furthermore, the OECTs can be powered by standard room lighting.

The authors demonstrated that their self-powered OPV–OECT sensing platform can detect biological signals (Fig. 1). They attached the platform to a person's finger and a gel electrode to the person's chest. Every heartbeat produced a voltage difference between the electrode and the platform, because of the movement of ions through the person's body. Such a difference is normally too small to be detected, but was measurable here owing to the high signal amplification achievable with the OECTs.

Park *et al.* found that, under a constant illumination from light-emitting diodes, the platform recorded clear heart-rate signals. The recording sensitivity was about three times higher than that of OECTs powered by conventional power sources[10]. This is because the absence of external power connections reduced signal fluctuations.

The authors validated the reliability of their approach *in vivo* by attaching the sensing platform to the exposed surface of a rat's heart. They successfully measured the animal's heart rate, proving the efficacy of the platform for monitoring biological signals.

© 2018 Springer Nature Limited. All rights reserved.

However, before Park and co-workers' OPV cells can be fully integrated into wearable devices, several optimizations are needed. The transmission of electronic signals from the platform is still based on conventional rigid silicon-based electronics that are powered by external sources. In addition, the OPV cells produce only low output power, so it will be challenging to drive sensing systems that are complicated or that have high power consumption.

Nevertheless, the OPV cells are a milestone in the production of ultrathin and highly efficient solar cells for self-powering applications. Moreover, the devices pave the way for the development of ultraflexible, stretchable and even healable[11,12] self-powered biosensors[13] for the precise, sensitive and continuous measurement of biological signals. ■

Shiming Zhang *is at the California NanoSystems Institute, University of California, Los Angeles, California 90095, USA.* **Fabio Cicoira** *is in the Department of Chemical Engineering, Polytechnique Montréal, Quebec H3T 1J4, Canada.*
e-mail: fabio.cicoira@polymtl.ca

1. Nathan, A. *et al. Proc. IEEE* **100**, 1486–1517 (2012).
2. Rogers, J. A., Someya, T. & Huang, Y. *Science* **327**, 1603–1607 (2010).
3. Park, S. *et al. Nature* **561**, 516–521 (2018).
4. Tang, C. W. *Appl. Phys. Lett.* **48**, 183–185 (1986).
5. Rivnay, J. *et al. Nature Rev. Mater.* **3**, 17086 (2018).
6. Cicoira, F. *et al. Adv. Mater.* **22**, 1012–1016 (2010).
7. Xia, Y. & Whitesides, G. M. *Angew. Chem. Int. Edn* **37**, 550–575 (1998).
8. Bowden, N., Brittain, S., Evans, A. G., Hutchinson, J. W. & Whitesides, G. M. *Nature* **393**, 146–149 (1998).
9. Zhang, S. *et al. Chem. Mater.* **29**, 3126–3132 (2017).
10. Campana, A., Cramer, T., Simon, D. T., Berggren, M. & Biscarini, F. *Adv. Mater.* **26**, 3874–3878 (2014).
11. Zhang, S. & Cicoira, F. *Adv. Mater.* **29**, 1703098 (2017).
12. Oh, J. Y. *et al. Nature* **539**, 411–415 (2016).
13. Huynh, T.-P. & Haick, H. *Adv. Mater.* https://doi.org/10.1002/adma.201802337 (2018).

CELL BIOLOGY

A core problem in nuclear assembly

Chromosomes can exist outside the nucleus in rupture-prone structures called micronuclei. It emerges that micronuclei are fragile because their outer layer lacks some nuclear-envelope components. SEE LETTER P.551

MATTHIAS SAMWER & DANIEL W. GERLICH

The genome sequencing of some cancer cells revealed the occurrence of large-scale DNA rearrangements that are characteristic of a single catastrophic event[1]. This type of genetic abnormality is termed chromothripsis. It occurs in many types of human cancer, and is thought to drive tumour formation[2]. Chromothripsis can arise as a result of cell division if individual chromosomes do not join the other chromosomes within the newly formed nuclei but instead form rupture-prone structures called micronuclei[3,4]. Liu *et al.*[5] report on page 551 that the nuclear-envelope structure that encloses a micronucleus lacks some of the components of the nuclear envelope that surrounds a nucleus. They propose a mechanism for how this might occur, and how it might account for micronuclear fragility.

In the cells of organisms such as animals and plants, the genome is stored in a nucleus. This is surrounded by a nuclear envelope[6] composed of lipid membranes, proteins, and multi-protein structures termed nuclear-pore complexes, which provide a route for the transport of materials between the nucleus and the cytoplasm. During cell division, the nuclear envelope disassembles, and the replicated chromosomes bind to a structure called the spindle, which is made of protein filaments called microtubules. The chromosomes align on the spindle at the centre of the cell, and the two copies of each chromosome then separate and move away from the middle of the cell. At a later stage of cell division termed telophase, a nuclear envelope reassembles around the chromosomes to form a nucleus in each daughter cell.

However, if any chromosomes lag behind the rest and are still in the middle of the dividing cell during telophase, they become isolated and can form a separate small nucleus called a micronucleus. Micronuclei often rupture spontaneously[7], resulting in substantial damage to DNA and chromothripsis[3,4]. The basis for micronuclear fragility was unclear.

Studying cell division in human cells grown *in vitro*, Liu *et al.* found that lagging chromosomes formed micronuclei that lacked certain components of the nuclear envelope, such as proteins from the nuclear-pore complex and proteins called B-type lamins that are necessary for nuclear-envelope integrity. Indeed, deficiencies in B-type lamins were previously shown to trigger micronuclear fragility[7]. Deficiencies in the formation of nuclear-pore complexes in the nuclear envelope that surrounds a micronucleus might hinder the entry of the proteins needed for successful repair of DNA damage and for stabilization of the nuclear envelope, which might explain why micronuclei are prone to rupture[7] and chromothripsis[3,4].

When normal nuclear assembly occurs around the end of cell division, part of the emerging nucleus is in close contact with the spindle, where the local density of microtubules is high. This part of the emerging nucleus is referred to as a region that contains 'core' nuclear-envelope components (Fig. 1). The part of the nucleus that emerges in areas with low microtubule density is known as the region that contains 'non-core' nuclear-envelope

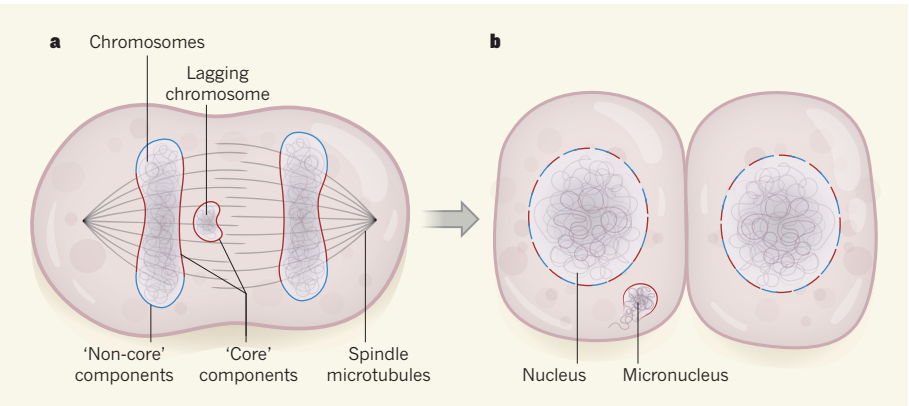


Figure 1 | The formation of a micronucleus. During cell division, chromosomes align in the middle of the cell on the spindle, which is made of protein filaments called microtubules. **a**, When chromosomes then move away from the middle of the dividing cell, a lagging chromosome might remain stranded in the centre of the cell, owing to a cell-division error. As cell division nears completion, the chromosomes become surrounded by what is termed nuclear-envelope material. 'Core' components of the nuclear-envelope material reassemble around chromosomes in spindle regions that have a high density of microtubules, and 'non-core' components reassemble in regions with a low density of microtubules. Liu *et al.*[5] analysed cell division in human cells grown *in vitro*. They report that the high density of microtubules surrounding lagging chromosomes prevents the binding of non-core components of the nuclear envelope. **b**, When nuclear-envelope assembly is completed, the core and non-core materials become interspersed in the nuclear envelope. When nuclear-envelope material reassembles around a lagging chromosome, the result is a rupture-prone structure termed a micronucleus that can undergo large-scale DNA damage and is linked to cancer[2–4].

© 2018 Springer Nature Limited. All rights reserved.

components. The authors found that nuclear-envelope components that incorporated efficiently into micronuclei were those components known[8] to localize to the core regions of reassembling nuclei. The nuclear-envelope components that were inefficiently incorporated or absent from micronuclei were those that are found on the non-core regions of reassembling nuclei.

Liu and colleagues investigated whether the failure of lagging chromosomes to recruit non-core nuclear-envelope components depends on the presence of the spindle. The authors treated cells undergoing cell division with drugs that dismantled the spindle. This treatment enabled both core and non-core nuclear-envelope components to assemble on lagging chromosomes that were forming micronuclei. The authors also conducted experiments in which they manipulated cells to displace some chromosomes away from the spindle to peripheral regions of a dividing cell. These displaced chromosomes formed micronuclei that had both core and non-core nuclear-envelope components. Such micronuclei did not rupture, revealing that non-core components are needed to prevent micronuclear fragility. The authors' results are consistent with a model in which the micronuclear assembly of nuclear-envelope material is perturbed in regions of the cell that have a high density of microtubules.

Cell division is a highly regulated process, and there are mechanisms to ensure its quality control. This raises the question of whether any such checkpoints exist in human cells to deal with lagging chromosomes.

In the fruit fly *Drosophila melanogaster*, there is a gradient of activity of the kinase enzyme Aurora B that is highest near the spindle in the centre of a dividing cell[9]. This enzymatic activity inhibits nuclear-envelope reassembly on lagging chromosomes that are within the region of high Aurora B activity. This mechanism for inhibiting nuclear-envelope reassembly might facilitate the reintegration of a lagging chromosome with the rest of the chromosomes that will form the new nucleus[9]. Indeed, when Aurora B is mislocalized away from the spindle in *Drosophila* cells[9], nuclear-envelope material, such as the non-core, nuclear-pore-complex protein Nup107, binds to lagging chromosomes with the same kinetics as to the main chromosome group.

However, when Liu *et al.* caused similar perturbations of Aurora B in human cells, this did not restore the recruitment of a non-core, nuclear-pore-complex protein (in this case, Nup133) to lagging chromosomes. The contrasting outcomes might reflect species-specific differences in sensitivity to Aurora B perturbations. Liu and colleagues found that core nuclear-envelope components accumulated on lagging chromosomes with normal kinetics and that drug-mediated inhibition of Aurora B at a late stage of cell division just before nuclear-envelope reassembly did not result in the recruitment of non-core, nuclear-envelope components to lagging chromosomes. These findings suggest that a dedicated chromosome-separation checkpoint is unlikely to exist in human cells. Instead, chromosome segregation to the daughter cells and nuclear-envelope assembly are not closely coordinated events. This could explain why lagging chromosomes often form micronuclei with deficient nuclear envelopes, rather than being incorporated into the nucleus.

How microtubules might limit the access of non-core, nuclear-envelope components to lagging chromosomes is unknown. The non-core and core components of the nuclear envelope both include membrane-bound proteins that redistribute during cell division to an organelle called the endoplasmic reticulum. One possibility is that these core and non-core proteins localize to different parts of the endoplasmic reticulum. Perhaps sheet-like domains of the organelle that are too large to enter spindle regions that have high microtubule density might contain non-core proteins, whereas the smaller tubular structures of the endoplasmic reticulum that might be able to enter the spindle regions could contain the core proteins. Alternatively, perhaps the localization of core and non-core nuclear-envelope proteins might be regulated by the microtubule-associated proteins that remove nuclear-envelope lipid membranes from chromosomes during an early stage of cell division[10]. It will be exciting to learn the outcome of studies to test these and other possible mechanisms, because the results might provide insights into general principles of how microtubules and other filament-like networks in the cell regulate the membrane-bound outer layer of organelles. ■

Matthias Samwer *is at Boehringer Ingelheim, 1120 Vienna, Austria.* **Daniel W. Gerlich** *is at the Institute of Molecular Biotechnology, Austrian Academy of Sciences, Vienna BioCenter, 1030 Vienna, Austria.*
e-mails: matthias.samwer@boehringer-ingelheim.com; daniel.gerlich@imba.oeaw.ac.at

1. Stephens, P. J. *et al. Cell* **144**, 27–40 (2011).
2. Ly, P. & Cleveland, D. W. *Trends Cell Biol.* **27**, 917–930 (2017).
3. Crasta, K. *et al. Nature* **482**, 53–58 (2012).
4. Zhang, C.-Z. *et al. Nature* **522**, 179–184 (2015).
5. Liu, S. *et al. Nature* **561**, 551–555 (2018).
6. Ungricht, R. & Kutay, U. *Nature Rev. Mol. Cell Biol.* **18**, 229–245 (2017).
7. Hatch, E. M., Fischer, A. H., Deerinck, T. J. & Hetzer, M. W. *Cell* **154**, 47–60 (2013).
8. Dechat, T. *et al. J. Cell Sci.* **117**, 6117–6128 (2004).
9. Afonso, O. *et al. Science* **345**, 332–336 (2014).
10. Schlaitz, A.-L., Thompson, J., Wong, C. C. L., Yates, J. R. & Heald, R. *Dev. Cell* **26**, 315–323 (2013).

This article was published online on 19 September 2018.

IMMUNOLOGY

Antibodies pose a double threat to HIV

Drug treatments for HIV infection require the long-term use of daily medication that can have toxic side effects. A pair of HIV-targeting antibodies might offer an alternative therapeutic approach. SEE ARTICLE P.479

NANCY L. HAIGWOOD

It is estimated that 36.9 million people worldwide are infected with HIV (go.nature.com/2pugvhd). Drug treatment can prevent this infection from developing into AIDS and becoming lethal, by controlling the level of virus in the bloodstream. However, this therapy doesn't eradicate HIV, and is associated with toxicity and conditions such as metabolic disorders[1]. As part of ongoing efforts to improve the treatment options available, clinical-trial results are now reported in two papers from the same research group — one in *Nature* by Mendoza *et al.*[2] (page 479) and the other by Bar-On *et al.*[3] in *Nature Medicine*. The studies assess the effects of a treatment for HIV that uses two antibodies to target a viral protein.

The current standard treatment for HIV infection is long-term use of antiretroviral therapy (ART), a daily regime of drugs that block steps needed for viral replication. If ART treatment ceases, virus becomes detectable in the bloodstream within days, a phenomenon termed viral rebound that can lead to progression towards AIDS. ART does not provide a cure because the drugs do not kill infected cells in the viral reservoir[4], which is established when virus inserts its genome into the genome of a host immune cell (Fig. 1).

ART drugs can prevent viral production from the part of the reservoir termed the active viral reservoir, in which viral replication occurs. However, these drugs cannot target the viral reservoir that exists in a 'dormant' state, termed the latent viral reservoir. A latent viral reservoir can exist in certain tissues[5]; here, virus production occurs at an extremely low level and might increase in the future. This 'hidden' reservoir can evade destruction by

© 2018 Springer Nature Limited. All rights reserved.

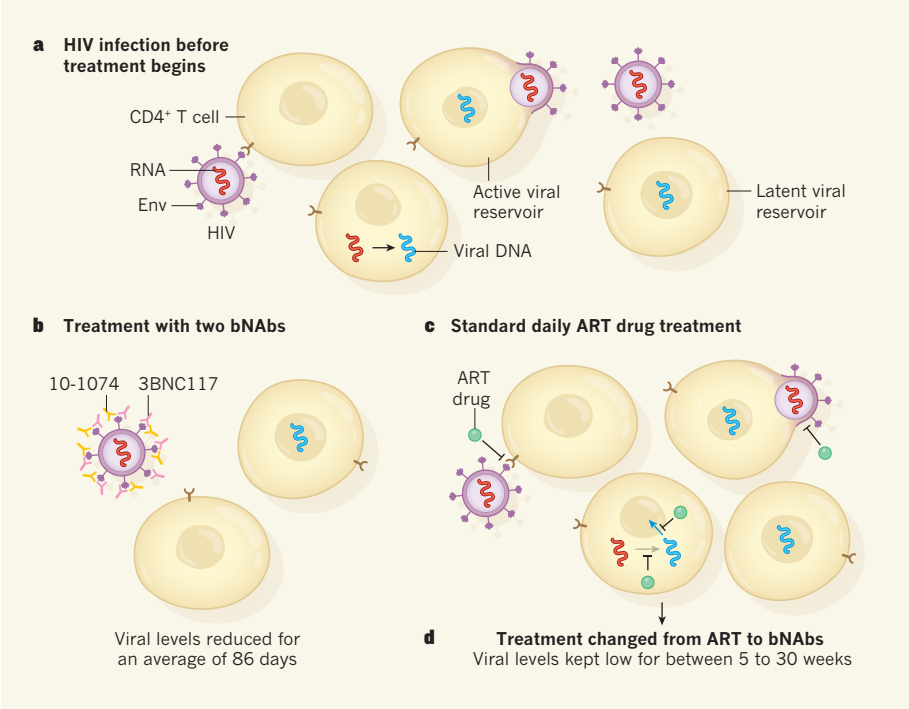


Figure 1 | Targeting HIV. **a**, When a person is infected with HIV, the protein Env on the viral surface can bind to receptors on immune cells called $CD4^+$ T cells. This interaction enables the virus to enter the cell, undergo DNA synthesis using the viral RNA template, and become inserted into the host-cell genome. Cells that are actively making virus using these inserted copies of viral DNA are called the active viral reservoir, and virus particles are released from such cells after viral replication. However, some cells that have viral DNA insertions might be in a 'dormant' state that does not actively produce virus and instead forms what is known as the latent viral reservoir; these cells might give rise to virus production in the future. **b**, Bar-On *et al.*[3] report the results of a clinical trial that tested whether the introduction of two antibodies, 3BNC117 and 10-1074, which are a type of antibody known as a broadly neutralizing antibody (bNAb), can lower the blood levels of HIV in people who haven't received HIV treatment. The two antibodies bind to separate sites on Env, and prevent the virus from binding and infecting immune cells. **c**, The standard treatment for HIV infection is known as antiretroviral therapy (ART), and consists of a daily dose of drugs that block steps in viral replication. **d**, Mendoza *et al.*[2] report a clinical trial that tested whether 3BNC117 and 10-1074 can lower virus levels in the bloodstream of people who temporarily stop receiving ART. The results of both studies are encouraging, indicating that the use of two bNAbs can lower virus levels for a time.

the immune system. If these cells could be targeted, more-effective treatments might be possible.

There has been interest in the potential use of HIV-targeting antibodies to control the viral level in the bloodstream. Such antibodies can bind to and block a protein on the viral surface called Env, which is needed for HIV entry into host immune cells called $CD4^+$ T cells. The discovery[6] of highly potent antibodies, termed broadly neutralizing antibodies (bNAbs), boosted this idea. These antibodies target regions of the Env molecule that are present in nearly all HIV strains. Individual bNAbs have already been tested[6–9] for use as clinical treatments for HIV. They are safe and well tolerated[10,11], and, in contrast to the daily dosage necessary for ART, administration is needed only every few weeks to maintain constant bNAb levels in the bloodstream and potentially in tissues.

Previous clinical studies[7–9] testing the individual effects of two bNAbs[12,13] — either 3BNC117 or VRC01 — found that the level of HIV in the bloodstream was initially suppressed for 6–10 weeks, but then viral rebound occurred, and the studies[7–9] reported the presence of antibody-resistant viral variants.

Mendoza *et al.* and Bar-On *et al.* carried out phase Ib clinical trials (small-scale trials to test the safety of a treatment) to investigate whether combining two bNAbs (3BNC117 and 10-1074) that target distinct sites on Env might decrease the probability of virus resistance occurring, and might control virus levels in HIV-infected people who did not receive ART during the trial period. Mendoza and colleagues assessed the effect of giving the two antibodies to 11 people who were temporarily stopping ART. Bar-On and colleagues analysed the effect of these two bNAbs on seven people who hadn't yet received ART.

Both studies reported an impressive reduction in bloodstream HIV levels compared with levels at the start of the trial. Viral rebound took many weeks to occur, and when it did, there was little or no evidence that viruses were resistant to both bNAbs. This suggests that using two antibodies might offer an alternative treatment option to ART, pending further studies.

Mendoza and colleagues observed that resistance to 10-1074 developed more rapidly than that to 3BNC117. This could be explained if there was a more rapid natural decline in the blood levels of 3BNC117, which had a shorter half-life than 10-1074. It effectively resulted in periods of treatment with a single type of bNAb rather than with the intended two. If the antibody dosage could be adjusted so that the half-lives of the two bNAbs were more closely matched, the development of treatment-resistant viral variants might be avoided. However, triple antibody combinations might be necessary to tackle the emergence of antibody resistance.

Maintained viral suppression from antibody treatment was evident in Mendoza and colleagues' study. The authors found that when the bNAbs were introduced 2 days before ART was stopped, and further doses given 3 and 6 weeks later, virus in the bloodstream was suppressed to undetectable levels for a median time of 21 weeks before viral rebound. The rebound timeframe ranged from 5 to 30 weeks, and 9 of the 11 trial participants maintained viral suppression without rebound for more than 15 weeks.

One way to determine whether the characteristics of viral infection change after bNAb therapy is to compare the diversity of the nucleotide sequences that encode the Env protein before and after treatment. Mendoza and colleagues found that the rebounding viruses after antibody treatment had low sequence diversity, whereas a high degree of diversity has been observed in viral rebound after ART treatment[14,15]. It is possible that this difference arises because the antibodies restrict the growth of viruses from the tissue viral reservoir, in addition to targeting actively replicating viruses in the bloodsteam, leading to a low level of sequence diversity.

How effectively might antibodies control virus levels without pretreatment with ART? In Bar-On and colleagues' study, the treatment suppressed virus levels in the bloodstream for an average of 86 days, about 60 days longer than previously observed[7] after treatment with just one HIV-targeting bNAb. This dual treatment approach in people who already had detectable levels of HIV in their bloodstream clearly limited the emergence of resistant viral variants, providing a notable advance on treatment with a single type of antibody. However, suppression was complete only in people who had low levels of virus in their bloodstream at the beginning of the trial.

The results provide cause for optimism that a cocktail of antibodies might provide substantial and durable viral control in the absence of ART. However, such enthusiasm must be tempered with caution. Bar-On and colleagues initially conducted laboratory tests to determine whether their study participants had a viral strain that would be sensitive to the bNAbs. In three individuals in whom the virus level rebounded most quickly, these antibody-sensitivity tests failed to detect viruses that had partial or complete pre-existing resistance to one or

© 2018 Springer Nature Limited. All rights reserved.

both bNAbs. This suggests that a more-sensitive test than that used by the authors will be needed to assess whether people are likely to respond to a particular bNAb combination.

In the war against HIV, there has been some controversy regarding the use of bNAbs as a treatment approach. Antibodies have been thought of as providing defences equivalent to a naval blockade, stopping HIV from initially infecting or preventing the infection from spreading. But once this barrier is breached and infection takes hold, it is assumed that T cells are the front line of immune defence against infected cells, with antibodies relegated to bystander status. However, models of infection with a simian–human chimaeric virus called SHIV in primates suggest that antibodies might have an active role in tackling infected cells.

In these primate models, human antibodies become widely distributed in tissues[16], and could be acting in a manner equivalent to minesweepers, by limiting the spread of infection. Not only are such antibodies distributed efficiently to tissues far from the site of antibody administration, but they can also help to destroy infected cells, at least in the early stages of an infection, and to prevent the establishment of the viral reservoir[16,17]. Once viral infection is established, the introduction of bNAbs that target Env can suppress the virus to undetectable levels in the bloodstream, as long as the antibodies persist there[18,19]. The role of these antibodies in tissues is not fully understood. When the antibody levels decay, viral rebound occurs and immune defences are mediated mainly by T cells[20].

Given the results in primates, is there any evidence that bNAbs can tackle the viral reservoir in human HIV infection? Mendoza and colleagues used an *in vitro* quantitative virus-outgrowth test to assess the viral-reservoir size. They found no significant differences in reservoir size before and after bNAb treatment. However, longer and more-effective bNAb treatment and larger study groups might be needed to establish definitively whether bNAbs affect the viral reservoir.

As strategies are sought to target the reservoir, it is encouraging to know that bNAbs might provide a useful weapon in the arsenal of viral-control tools. Moreover, if it turns out that a cocktail of bNAbs could act as a potential temporary replacement for ART, this type of 'drug holiday' might give people time to recover from ART-induced toxicity. ■

Nancy L. Haigwood *is in the Division of Pathobiology & Immunology, Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, Oregon 97006, USA.*
e-mail: haigwoon@ohsu.edu

1. Paula, A. A., Falcão, M. C. N. & Pacheco, A. G. *AIDS Res. Ther.* **10**, 32 (2013).
2. Mendoza, P. *et al. Nature* **561**, 479–484 (2018).
3. Bar-On, Y. *et al. Nature Med.* https://doi.org/10.1038/s41591-018-0186-4 (2018).
4. Sengupta, S. & Siliciano R. F. *Immunity* **48**, 872–895 (2018).
5. Boritz, E. A. *et al. Cell* **166**, 1004–1015 (2016).
6. Pegu, A., Hessell, A. J., Mascola, J. R. & Haigwood, N. L. *Immunol. Rev.* **275**, 296–312 (2017).
7. Caskey, M. et al. *Nature* **522**, 487–491 (2015).
8. Bar, K. J. *et al. N. Engl. J. Med.* **375**, 2037–2050 (2016).
9. Scheid, J. F. *et al. Nature* **535**, 556–560 (2016).
10. Ledgerwood, J. E. *et al. Clin. Exp. Immunol.* **182**, 289–301 (2015).
11. Mayer, K. H. *et al. PLoS Med.* **14**, e1002435 (2017).
12. Scheid, J. F. *et al. Science* **333**, 1633–1637 (2011).
13. Zhou, T. *et al. Science* **329**, 811–817 (2010).
14. Rothenberger, M. K. *et al. Proc. Natl Acad. Sci. USA* **112**, E1126–E1134 (2015).
15. Kearney, M. F. *et al. PLoS Pathog.* **10**, e1004010 (2014).
16. Hessell, A. J. *et al. Nature Med.* **22**, 362–368 (2016).
17. Liu, J. *et al. Science* **353**, 1045–1049 (2016).
18. Shingai, M. *et al. Nature* **503**, 277–280 (2013).
19. Barouch, D. H. *et al. Nature* **503**, 224–228 (2013).
20. Nishimura, Y. *et al. Nature* **543**, 559–563 (2017).

ENGINEERING

Complex motion guided without external control

Mechanical structures have been made that exhibit self-guided, multi-step sequences of shape changes in response to an applied force. Such structures could have applications in flexible electronics and soft robotics. SEE LETTER P.512

LARRY L. HOWELL

It is reasonable to expect that to guide a mechanical structure through a complex, predetermined series of shape changes, electronic controls would be needed. In many cases, motors move the structure, sensors detect position and electronic devices called microcontrollers execute algorithms to provide external control. Achieving complex motion using only the structure's geometry might seem unrealistic, but, if possible, it would have many advantages. In particular, it would simplify the manufacture of mechanical structures and enable them to be readily produced in various sizes. On page 512, Coulais *et al.*[1] report an approach for designing materials that follow a prescribed motion and that then reconfigure at specified points to enter another phase of motion. This movement is guided only by the material's geometry, which eliminates the need for external sensing and control.

Engineered structures known as mechanical metamaterials are designed to behave in ways not common in nature[2–4]. They are usually made by combining building blocks (unit cells) into patterns, to produce a combined structure that can behave in ways not possible using a solid material. Advances in materials, manufacturing, computational capabilities and mechanical-design approaches are enabling designs that were not previously possible, such as the mechanical metamaterials described by Coulais and colleagues.

The authors used thin, flexible parts, strategically combined with thick, stiff parts, to achieve the desired structural motion. They refer to the movement of these parts as soft deformation, because it is caused by the bending of the flexible parts, rather than by mechanical hinges. This type of motion is not new, of course, because it can be seen extensively in nature. Consider, for example, the movement of an elephant's trunk, your beating heart or a butterfly's flapping wings — in each case, motion results from the bending of flexible parts.

Designs in which a combination of flexible and stiff components enables a specified motion are often referred to as compliant mechanisms[5]. The use of these mechanisms in engineered materials offers several advantages over conventional design methods, such as the ability to achieve complex motion from a single part rather than from an assembly of pieces connected by hinges. Coulais *et al.* used compliant mechanisms to realize surprisingly sophisticated multi-step motions (Fig. 1). One clever aspect of their approach is the use of parts that have different degrees of flexibility, so that, in response to an external force, some parts deform before others do.

Coulais and colleagues also used what they call self-contact, whereby a stiff part is intentionally brought into contact with another stiff part to prevent further motion between the two. This 'contact-aided' mechanism[6,7] is similar to a door-stop that prevents a door from colliding with a wall. In the authors' approach, bending of the most flexible part results in self-contact, which ends the first step of a multi-step pathway (Fig. 1a). The most flexible part then becomes inactive, and the second step begins (Fig. 1b). The next-most flexible part becomes the main source of deformation, and this part bends until the next occurrence of self-contact.

The authors produced unit cells from these flexible and stiff components, and combined the unit cells into patterns to achieve the desired motions. In one demonstration, they considered unit cells consisting of a cross-shaped arrangement of five equally sized squares, which were

© 2018 Springer Nature Limited. All rights reserved.

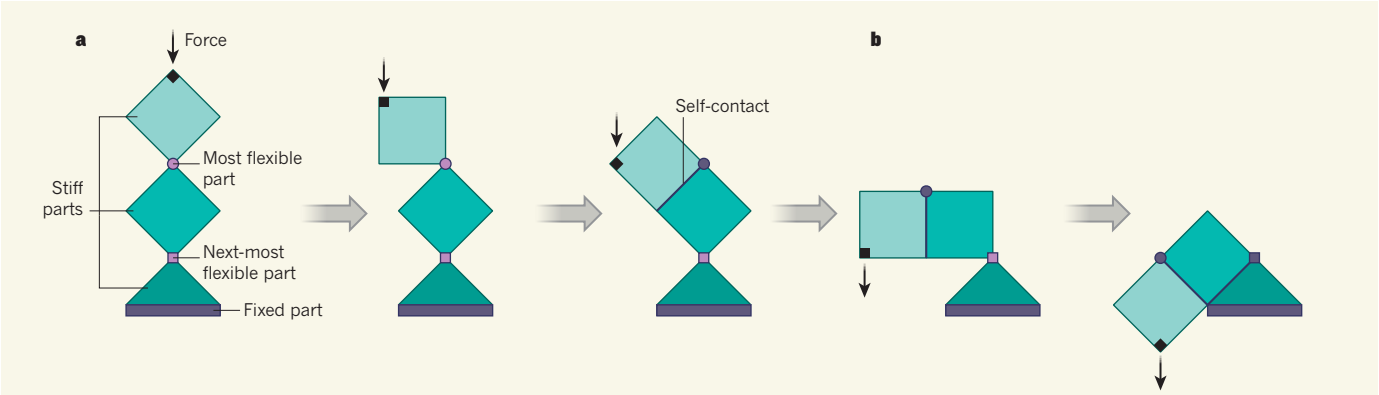


Figure 1 | Self-guided, multi-step motions of a mechanical structure. Coulais *et al.*[1] demonstrate an approach for designing materials that follow a multi-step pathway consisting of a series of shape changes, without needing external control. The materials are composed of building blocks (unit cells), and a subset of a unit cell is shown here. **a**, The subset consists of stiff parts that are connected by flexible parts, and a part that is fixed. The flexible parts have different degrees of flexibility. The authors apply an external force to the structure (at the point marked by the black square), which causes the most flexible part to bend. This bending moves one of the stiff parts until it connects with another stiff part — a process known as self-contact. The subset reconfigures such that the most flexible part becomes inactive and the connected parts act as a single stiff part. The first step of the multi-step pathway is complete. **b**, The next-most flexible part then bends until another self-contact occurs, which prevents further motion. The desired configuration is reached, and the second step of the pathway is concluded.

connected by flexible parts (see Fig. 1a of the paper[1]). In the first step of the multi-step pathway, each unit cell rotated to produce a structure of differently sized squares. In the second step, each of these squares rotated to fill an assigned space.

Coulais and colleagues illustrated the concepts associated with their approach using only a few specific examples, but their work provides guidelines that will enable the design of other shape-changing metamaterials. The authors' unit cells can be configured in many different ways, resulting in different motions. For example, two structures that have nearly identical topologies can have contrasting pathways (see Fig. 1a,b of the paper[1]). Even more configurations are made possible by producing different arrays of the unit cells, or by using the principles discussed in the paper to make other unit cells and arrays.

The approach does have some limitations. For instance, the flexible parts are relatively small and undergo 90° of rotation, and few materials can withstand such large strains without breaking. Even materials that do not break might still become permanently deformed, making the process irreversible — and therefore unsuitable for many applications. The authors used 10-millimetre-thick silicon rubber for the flexible parts in their experiments. Although the use of this material is valuable for verifying and demonstrating the approach, issues will probably arise when smaller parts are required, because the material choice will be limited.

Coulais and colleagues' results open up many avenues for future research. A natural next step would be to see what other multi-step motions can be realized using the approach. Researchers could also study how the designs vibrate when they are shaken, to see whether they still undergo the intended movements, or whether they follow a different pathway.

In addition, it would be useful to explore whether switch-type mechanisms[8,9] can be used to snap the structures into desired configurations, replacing the self-contact function and perhaps eliminating the need for an external force to keep the structures in key intermediate positions. Finally, researchers could investigate how the principles described in the current work could be used to produce sophisticated motions beyond multi-step pathways, and how the approach might be applied to the micrometre and nanometre scales. ■

Larry L. Howell *is in the Department of Mechanical Engineering, Brigham Young University, Provo, Utah 84602, USA.*
e-mail: lhowell@byu.edu

1. Coulais, C., Sabbadini, A., Vink, F. & van Hecke, M. *Nature* **561**, 512–515 (2018).
2. Overvelde, J. T. B., Weaver, J. C., Hoberman, C. & Bertoldi, K. *Nature* **541**, 347–352 (2017).
3. Reis, P. M., Jaeger, H. M. & van Hecke, M. *Extreme Mech. Lett.* **5**, 25–29 (2015).
4. Haghpanah, B., Salari-Sharif, L., Pourrajab, P., Hopkins, J. & Valdevit, L. *Adv. Mater.* **28**, 7915–7920 (2016).
5. Howell, L. L., Magleby, S. P. & Olsen, B. M. *Handbook of Compliant Mechanisms* (Wiley, 2013).
6. Hopkins, J. Shaw, L., Dotson, M., Chizari, S. & Song, Y. *Bull. Am. Phys. Soc.* abstr. F17.00004 (2018).
7. Mankame, N. D. & Ananthasuresh, G. K. *Comput. Struct.* **82**, 1267–1290 (2004).
8. Silverberg, J. L. *et al. Nature Mater.* **14**, 389–393 (2015).
9. Waitukaitis, S., Menaut, R., Chen, B. G. & van Hecke, M. *Phys. Rev. Lett.* **114**, 055503 (2015).

COMPUTATIONAL BIOCHEMISTRY

Designer proteins activate fluorescence

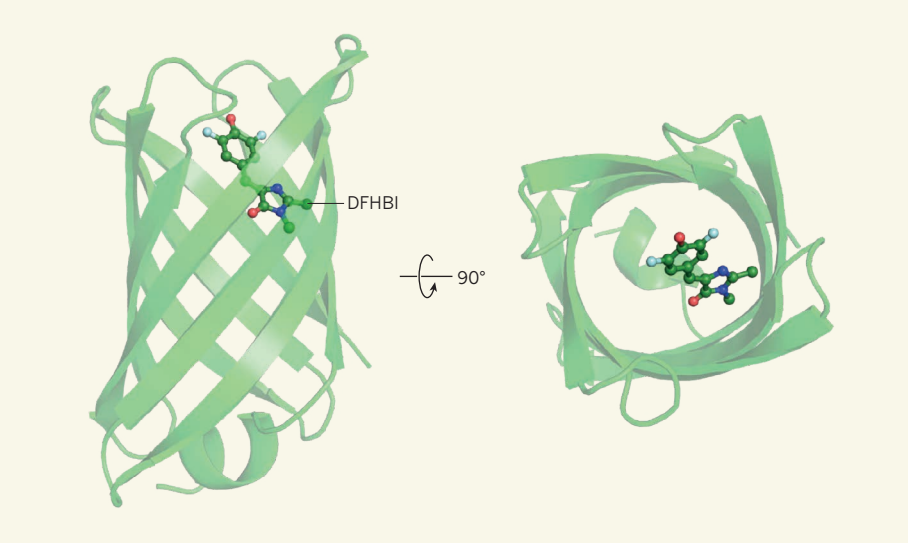
A computational method has been devised that allows a structural motif found in proteins, known as a β-barrel, to be designed to bind specifically to any small molecule, opening the door to biotechnological applications. SEE ARTICLE P.485

ROBERTO A. CHICA

Proteins are the molecular machines of life: they carry out the complex molecular processes required by cells with unrivalled accuracy and efficiency. Many of these processes depend on proteins having the ability to bind specifically to a given small molecule. If we could make proteins from scratch to bind any desired target molecule, it would open the door to a wide range of biotechnological applications that are not currently possible using natural proteins. On page 485, Dou *et al.*[1] describe a computational method for designing proteins tailored to bind a small molecule of interest, and use it to make 'fluorescence-activating' proteins — biotechnological tools that have potential applications in biomedical research.

The functions of proteins are dictated by

© 2018 Springer Nature Limited. All rights reserved.



ROBERTO A. CHICA

Figure 1 | Binding by design. Dou *et al.*[1] report a computational method for designing protein structures (known as β-barrels) that can bind any target molecule. They have used it to make small proteins that bind the molecule DFHBI, which fluoresces when bound. The crystal structure of one of the proteins in complex with DFHBI is shown here as a ribbon structure.

the specific three-dimensional structures into which they fold. One such structure is the β-barrel: a pleated β-sheet rolled into a cylindrical structure. This is frequently found in natural proteins whose function is to bind small ligand molecules[2]. So far, several challenges have prevented the *de novo* design of β-barrel proteins that can bind target small molecules. One issue is that the cylindrical shape of β-barrels requires the two edges of the original β-sheet to come together to form the base of the cylinder, a process that is complicated by the tendency of β-sheets to form intermolecular interactions that generate protein aggregates. Another is that these cylinders must be open at one end to allow the binding of a small molecule inside the barrel, which can cause substantial destabilization of the protein's folded structure. Thus, the design of ligand-binding β-barrel proteins requires a delicate balance between stability and ligand-binding activity.

Designing a cavity that is geometrically and chemically complementary to the small molecule of interest is just as challenging. This objective is complicated by the huge number of positions and orientations in space that the small molecule could adopt, and by the difficulty of identifying the exact combination of amino-acid residues that will make the necessary interactions with the ligand for specific binding. The latter point is particularly challenging given that there are some 10^{143} possible amino-acid sequences for proteins made up of 110 amino acids (as Dou and colleagues' β-barrel is). Thus, accurate and efficient computational procedures are required to find optimal solutions from such an astronomically large number of possibilities.

Dou *et al.* provide a solution to these problems. They started by devising general principles for designing stably folded β-barrel proteins of a predetermined size. They discovered that structural irregularities known as glycine kinks and β-bulges must be introduced into β-barrels to alleviate molecular strain and to maintain the continuous pattern of hydrogen bonds needed to form the cylindrical structure. Using their approach, they built computational models of 500 possible β-barrel 'backbones' and performed calculations to identify amino-acid sequences that would stabilize each backbone. The four designs that were predicted to be most stable were synthesized, and the authors observed that one of these folded into a monomeric β-barrel. Strikingly, the computationally predicted model and the experimentally determined structure were very similar, and the designed protein was highly stable.

The authors went on to design a β-barrel protein that binds DFHBI, a small molecule that fluoresces on binding. They used an innovative computational procedure that models a large number of poses (positions and orientations in space) of a small molecule and how they dock into a protein binding site while simultaneously optimizing protein sequences for binding activity. Through this approach, the authors generated 56 β-barrel designs that were predicted to fold stably and to bind tightly to DFHBI. Of these, 20 were found to be monomeric and soluble when tested experimentally, and exhibited some of the characteristic properties of β-sheets — indicating that they were potentially β-barrels. Two of those proteins bound DFHBI with moderate binding affinities (in the range of 13–50 micromolar).

Dou and colleagues next used an iterative procedure to further improve the binding affinity of their proteins. This included steps in which the X-ray crystal structures of two protein variants were used to guide additional rounds of computational design, and in which every amino-acid residue of another variant was systematically mutated to find changes that improved binding affinity. In this way, they identified three "mini-fluorescence-activating proteins" (Fig. 1) that bind DFHBI with submicromolar affinity and enhance its fluorescence both *in vitro* and *in vivo*. These designer proteins could be used to monitor gene expression and to track proteins in cells, as well as in biosensors that detect the presence of chemicals.

The development and application of this computational method for designing β-barrel proteins that bind small molecules is the first demonstration of the *de novo* design of both protein fold and function, a milestone in the field. Previous computational designs of ligand-binding proteins relied on building a binding cavity into a protein template found in nature[3], or one that had previously been created in the laboratory[4]. By contrast, Dou and co-workers have designed a β-barrel protein that has a shape distinct from those found in nature, and constructed a binding pocket that is specifically tailored to a target small molecule.

As noted earlier, the authors' initial designs needed further optimization to identify proteins that have sufficiently high binding affinities for potential applications. More-accurate predictions of protein structures are needed to eliminate the need for such fine-tuning. One way of achieving this might come from recognizing that proteins are not rigid molecules that adopt a single predominant structure — like all machines, proteins need to move to accomplish their tasks with high efficiency[5,6]. Indeed, ligand binding is often the trigger that causes a protein receptor to undergo a structural change enabling the transmission of a biological signal[7]. Computational methods for the rational design of proteins that undergo particular structural changes have recently been developed[8]. If these could be combined with Dou and colleagues' approach, it might be possible to access more-complex protein functions than were previously possible, opening the door to the on-demand creation of protein-based molecular machines. ■

Roberto A. Chica *is in the Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.*
e-mail: rchica@uottawa.ca

1. Dou, J. *et al. Nature* **561**, 485–491 (2018).
2. LaLonde, J. M., Bernlohr, D. A. & Banaszak, L. J. *FASEB J.* **8**, 1240–1247 (1994).
3. Tinberg, C. E. *et al. Nature* **501**, 212–216 (2013).
4. Polizzi, N. F. *et al. Nature Chem.* **9**, 1157–1164 (2017).
5. Kerns, S. J. *et al. Nature Struct. Mol. Biol.* **22**, 124–131 (2015).
6. Catterall, W. A., Wisedchaisri, G. & Zheng, N. *Nature Chem. Biol.* **13**, 455–463 (2017).
7. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. *Chem. Rev.* **117**, 139–155 (2017).
8. Davey, J. A., Damry, A. M., Goto, N. K. & Chica, R. A. *Nature Chem. Biol.* **13**, 1280–1285 (2017).

This article was published online on 12 September 2018.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0497-0

Population dynamics of normal human blood inferred from somatic mutations

Henry Lee-Six[1], Nina Friesgaard Øbro[2], Mairi S. Shepherd[2], Sebastian Grossmann[1], Kevin Dawson[1], Miriam Belmonte[2], Robert J. Osborne[1], Brian J. P. Huntly[2], Inigo Martincorena[1], Elizabeth Anderson[1], Laura O'Neill[1], Michael R. Stratton[1], Elisa Laurenti[2], Anthony R. Green[2,3]*, David G. Kent[2,3]* & Peter J. Campbell[1,3]*

Haematopoietic stem cells drive blood production, but their population size and lifetime dynamics have not been quantified directly in humans. Here we identified 129,582 spontaneous, genome-wide somatic mutations in 140 single-cell-derived haematopoietic stem and progenitor colonies from a healthy 59-year-old man and applied population-genetics approaches to reconstruct clonal dynamics. Cell divisions from early embryogenesis were evident in the phylogenetic tree; all blood cells were derived from a common ancestor that preceded gastrulation. The size of the stem cell population grew steadily in early life, reaching a stable plateau by adolescence. We estimate the numbers of haematopoietic stem cells that are actively making white blood cells at any one time to be in the range of 50,000–200,000. We observed adult haematopoietic stem cell clones that generate multilineage outputs, including granulocytes and B lymphocytes. Harnessing naturally occurring mutations to report the clonal architecture of an organ enables the high-resolution reconstruction of somatic cell dynamics in humans.

Human haematopoiesis balances the production and destruction of hundreds of billions of specialized blood cells every day. This process relies on a multilayered hierarchy of progressively more differentiated and more populous cells, at the top of which sits the pool of stem cells. First described functionally in the 1960s[1,2], haematopoietic stem cells are defined by their ability to establish long-term, stable contributions to multiple lineages of blood cells, including myeloid, T and B cells. The numbers and dynamics of stem cells in homeostatic human haematopoiesis remain poorly defined, despite their routine use in therapeutic transplantation for haematological diseases.

Historical studies in animals quantified haematopoiesis either by labelling cells in vitro and transplanting them into a recipient animal[3–6] or by modelling X chromosome inactivation patterns[7]. More recently, studies tracking the clonal contributions of cells labelled directly in vivo[8–10] have suggested that long-term homeostatic haematopoiesis is driven by many thousands of cells that do not function as classical stem cells in transplantation assays[11].

Approaches to measuring the dynamics and potential of stem cells in humans have been less direct. Studies of X chromosome inactivation patterns in haematological malignancies have demonstrated their clonal origin in stem cells that had multilineage potential[12,13]. In patients who received gene therapy, hundreds to thousands of clones contribute to lymphoid and myeloid lineages more than a year after transplantation[14,15]. Studies of unperturbed haematopoiesis in humans have relied on ex vivo cellular assays[16] or modelling of telomere lengths[17] and X chromosome inactivation patterns[18]. These analyses have suggested that the numbers of stem cells increase through childhood and adolescence, reaching a plateau in adulthood, with some shift in lineage potential.

Spontaneous mutations used to measure haematopoiesis

Mutations accumulate in somatic cells throughout life[19,20]. A mutation arising in a cell is inherited by its descendant cells, a feature that has enabled the reconstruction of clonal structures in cancer[21] and normal development[22,23]. In normal blood stem cells, the burden of somatic mutations increases linearly with age[20], suggesting that they represent an accurate molecular clock.

We hypothesized that spontaneous somatic mutations could act as clonal markers, thus enabling the quantification of the number, activity and longevity of human blood stem cells during normal haematopoiesis. Analogous to capture–recapture experiments in ecology, our design followed two phases (Fig. 1). First, in the 'capture' phase, we isolated single haematopoietic stem and progenitor cells[24] from a bone marrow aspirate and peripheral blood draw from a 59-year-old male with normal blood counts and no past history of blood disorders (Extended Data Fig. 1). These were expanded in single-cell liquid cultures or colony-forming cell assays. We performed whole-genome sequencing on 198 colonies, each to around 15× depth (Supplementary Table 1), and identified somatic mutations. Second, in the 'recapture' phase, we isolated bulk populations of mature peripheral blood cells from the same individual: granulocytes at three time points after the bone marrow aspirate, together with B and T lymphocytes, both from one time point. We performed deep targeted sequencing on these bulk populations for mutations discovered in the capture phase.

Bringing together stem cell biology and population genetics creates a risk of lexical confusion. We reserve the term 'clone' for the in vivo descendants of a single ancestral cell; and use 'colony' to describe the cells derived in vitro from a single stem or progenitor cell. We use 'lineage' to denote a specific functional group of blood cells, such as granulocytes; and 'line-of-descent' for the set of cells that are direct antecedents/descendants of the cell in question (see Supplementary Information).

Mutation burden and spectrum

In total, 140 colonies had variant allele fractions that were distributed around 50%, confirming that they did in fact derive from a single cell, whereas 58 of the colonies had lower allele fractions (Extended Data Fig. 2 and Supplementary Table 1); this was probably caused by colonies

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. [2]Wellcome-MRC Cambridge Stem Cell Institute and Department of Haematology, University of Cambridge, Cambridge, UK. [3]These authors jointly supervised this work: Anthony R. Green, David G. Kent, Peter J. Campbell. *e-mail: arg1000@cam.ac.uk; dgk23@cam.ac.uk; pc8@sanger.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

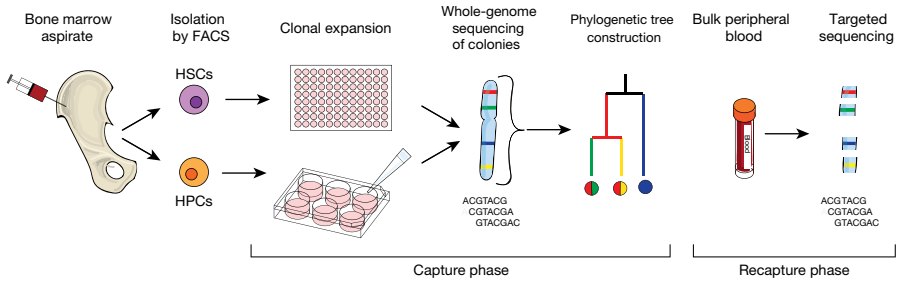


Fig. 1 | Experimental design. The experiment proceeded in two phases: a capture phase, in which single haematopoietic stem and progenitor cells were expanded in vitro and were analysed using whole-genome sequencing, and a recapture phase, in which bulk populations of differentiated cells were deep-sequenced for mutations that were identified in the capture phase. FACS, fluorescence-activated cell sorting; HPCs, haematopoietic progenitor cells; HSCs, haematopoietic stem cells.

growing into each other in the methylcellulose. These polyclonal colonies were excluded from further analyses. It proved more difficult to derive clonal colonies from some progenitor types than others, such that our final set of 140 colonies was composed of 89 immunophenotypic haematopoietic stem cells, 38 megakaryocyte–erythrocyte progenitors, eight granulocyte–macrophage progenitors and five common myeloid progenitors.

We assessed whether variants were acquired during in vitro expansion. Any mutation on the X chromosome in the original colony-forming cell should be present at allele fractions near 100%. The mean percentage of X chromosome mutations found on <50% reads was only 5.6% per colony, suggesting that variants acquired in vitro were infrequent.

Mutation burden was consistent across colonies, with a mean of 1,023 substitutions (range, 815–1,210) and 20 small insertion/deletion (indel) events (range 2–37) (Extended Data Fig. 3). No somatically acquired structural variants were identified. The spectrum of mutations was dominated by C > T and T > C transitions, as described in myeloid cancers[20] and age-related clonal haematopoiesis[25] (Extended Data Fig. 4).

Driver mutations in myeloid cancer genes occur in some older individuals with normal blood counts[26–28]. In the colonies from our subject, we found no known myeloid driver mutations. Furthermore, the ratio of non-synonymous to synonymous mutations, a metric that can detect positive or negative clonal selection across genes[29], was exactly as expected for the background mutation spectrum ($dN/dS = 1.001$; 95% confidence interval, 0.889–1.127; with $dN/dS = 1$ representing neutrality). Finally, driver mutation hotspots were included in the bait set for the recapture phase (Supplementary Table 2), and no such drivers were detected. Thus, haematopoietic cells in this subject have undergone selectively neutral accumulation of somatic mutations.

Somatic mutations acquired in embryo development

To explore clonal relationships among the 140 colonies, we constructed a phylogenetic tree (Fig. 2 and Extended Data Figs. 5, 6). Of the 129,582 somatic substitutions placed on the tree, 8,676 were seen in more than one colony. At the top of the tree, two mutations completely partitioned the colonies, one found in 52 and the other in 88 colonies, with every colony possessing one or other mutation, and no colony possessing both (Fig. 2b). These same two mutations were found in a buccal swab taken from the patient, and at clonal contributions that suggested the same two thirds to one third split. This indicates that the most recent common ancestor of all blood cells in this subject was also the most recent common ancestor of buccal epithelial cells. Because blood derives from the mesoderm and the buccal epithelium from the ectoderm, this common ancestor must have predated gastrulation. It is likely that this most recent common ancestor of both blood and buccal epithelium was in fact the fertilized egg: the two observed mutations would have occurred during its first cell division, one to each daughter cell. These two daughter cells then contributed unequally to adult somatic tissues, as previously observed[22,30,31].

After the first division, a cascade of further mutations provides ever finer partitions of the colonies (Fig. 2b), consistent with data in mice that adult haematopoiesis is a mosaic of embryonic clones[32]. By 10 mutations of molecular time, 33 lines-of-descent were created, which required at least 5 generations of cell doublings. Embryonic lineages that are lost or unobserved would suggest more than 5 generations, so average mutation rates in early embryogenesis could be around 2 per division at most. Of the 32 cell divisions by 10 mutations in molecular time, at least 10 were associated with no mutations (Fig. 2c), noting that unobserved embryonic lineages can convert polytomies (multiway splits) to dichotomies. This provides an estimate of the mutation rate of 1.2 per division (Supplementary Methods). Thus, mutation rates in embryonic cells that ultimately contribute to somatogenesis fall in the range of 1–2 per cell division, similar to estimates from human neural progenitor clones[23] and de novo germline mutations[33].

Clonal relationships of haematopoietic cells

Clonal relationships among the 140 colonies were evident beyond the early embryonic mutations as a series of branch points scattered down the vertical axis of the phylogenetic tree (Fig. 2a). Mutations shared between two colonies indicate that they derive from a common ancestor; the ratio of shared to unique mutations is a measure of the age at which the two lines-of-descent split. Long lines-of-descent with few branches suggest that we have sequenced only a small fraction of the stem cell pool.

The distribution of different cell types across the tree provides information on population stratification among haematopoietic stem and progenitor cells. Stem cells from the bone marrow aspirate, taken from the right iliac crest, were no more clustered together on the phylogenetic tree than those derived from peripheral blood ($P = 0.14$; Fig. 2 and Extended Data Fig. 3a). This suggests that stem cells recirculate and redistribute sufficiently often that the population in the iliac crest is a random sample of the whole-body stem cell pool. Similarly, progenitors were not more clustered on the tree than stem cells ($P = 0.12$; Extended Data Fig. 3b, c), suggesting that the progenitors that we sequenced are not drawn from a more restricted set of historic lines-of-descent than the stem cells.

The interspersed distribution of stem and progenitor cells on the tree and relative scarcity of recent branch points indicates that the phylogeny is dominated by events that occurred in stem cells. As a progenitor is short-lived, only the most recent mutations in its line-of-descent will have occurred while it was a progenitor. Therefore, a branch point hundreds of mutations ago represents an ancient symmetrical division in which one stem cell gave rise to two stem cells, since descendants of both daughters persist decades later as haematopoietic cells. In traditional experimental models, long-term self-renewal of haematopoietic stem cells is established prospectively (through serial transplantation assays), whereas here we infer it retrospectively through the decades-long persistence of lines-of-descent from an ancestral stem cell division. Therefore, for the analyses that follow, which rely

© 2018 Springer Nature Limited. All rights reserved.

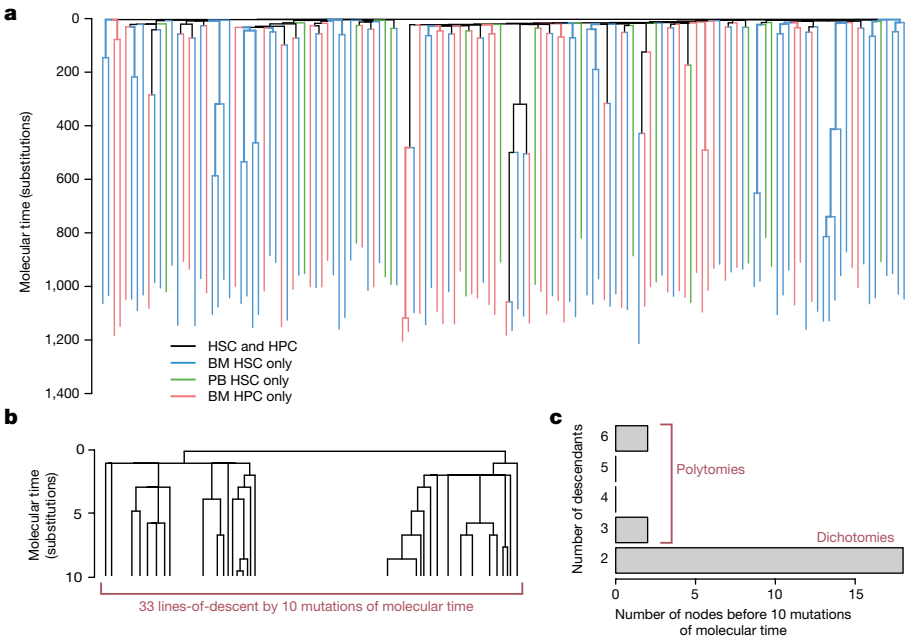


Fig. 2 | The phylogeny of cells, showing the relationship between cell types and embryological cell divisions. a, Phylogeny of 140 single haematopoietic stem and progenitor cells showing the relationship between cell types. At each tip of the tree is a colony. Branches connect colonies to each other to form a family tree. Branch lengths are proportional to the number of somatic mutations. Branches are coloured according to the phenotype of their descendants. Branches ancestral to haematopoietic progenitor cells (HPCs) are coloured red; branches ancestral to bone marrow-derived haematopoietic stem cells (BM HSCs) are blue; branches ancestral to peripheral blood-derived haematopoietic stem cells (PB HSCs) are green; branches ancestral to both stem and progenitor cells are coloured black. **b**, The same phylogeny as in **a**, but showing only the first 10 mutations of molecular time. **c**, The number of descendants of each node for the first 10 mutations of molecular time, used to estimate the embryonic mutation rate.

on mutations distributed across the tree, we need not distinguish whether the colony that was sequenced is derived from a stem or progenitor cell.

Lifetime trajectory of the population size of stem cells

The relative timings of branch points, or more formally 'coalescences', in the phylogenetic tree inform on historical population dynamics. In brief, under neutral drift, the rate at which lines-of-descent coalesce is related to both the effective population size and the generation time (here, the time between symmetrical stem cell divisions). Methods to infer historic population dynamics from coalescences[34] were applied to the tree built from the blood cells of our subject. This revealed a rapid population expansion of haematopoietic stem cells during early life, reaching a relatively stable plateau by late childhood or early adolescence (Fig. 3), consistent with previous inferences[17,18]. The stable population size during adulthood suggests that symmetric self-renewal divisions, during which one stem cell divides into two stem cells, are balanced by stem cell death, senescence and symmetric divisions into committed progenitors.

The same broad coalescence structure of the phylogenetic tree that arises with a given population size and number of generations can be generated by a population ten times larger going through ten times as many generations. Therefore, from the structure of the tree alone, without knowing how many generations the cells have been through, we cannot directly estimate the absolute number of stem cells. We therefore performed deep-sequencing of blood cells in the recapture phase of our study.

Estimates of stem cell number and generation time

We designed a hybridization bait set for 7,116 mutations that were identified in the colonies and assigned to the phylogenetic tree, choosing 6,317 mutations that were shared by more than one colony and 799 that were unique to single colonies. We performed targeted sequencing of three peripheral blood granulocyte samples taken 4 months (mean coverage, 776×), 9 months (mean coverage, 4,669×) and 14 months (mean coverage, 268×) after the bone marrow aspirate, together with control cord blood from two individuals (mean coverage, 5,305×) (Supplementary Table 3). We used Bayesian generalized linear mixed models to estimate the fraction of reads that were derived from true mutant alleles versus sequencing errors. Estimated allele fractions were stable across the three time points and steadily decreased down individual lines-of-descent (Extended Data Fig. 8).

Three observations emerge (Fig. 4). First, the majority of mutations in the bait set were not detectable (horizontal grey ticks in Fig. 4). This suggests that the number of active stem cells in our subject must be much higher than a few thousand, since our detection threshold was

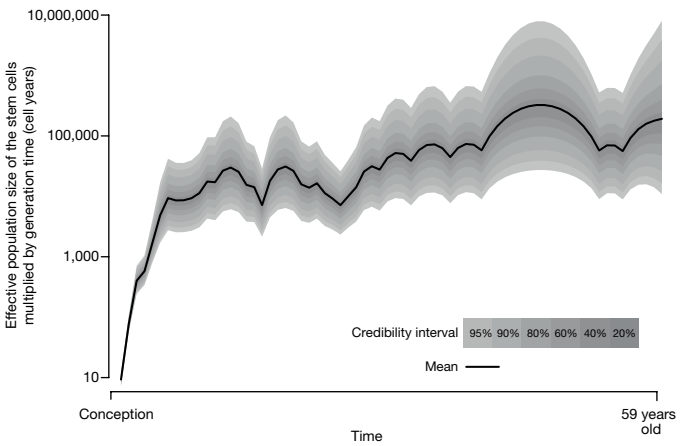


Fig. 3 | Population size trajectory of the stem cell pool. Phylodynamic methods reveal changes in the effective population size of stem cells over life based on the timing of coalescences (branch points) in our observed phylogeny. Shading illustrates different credibility intervals. The *y* axis is shown in units of 'population size multiplied by generation time' (cell years), because the same distribution of coalescences can be generated from a population of 10 times the size with 10 times as many generations.

© 2018 Springer Nature Limited. All rights reserved.

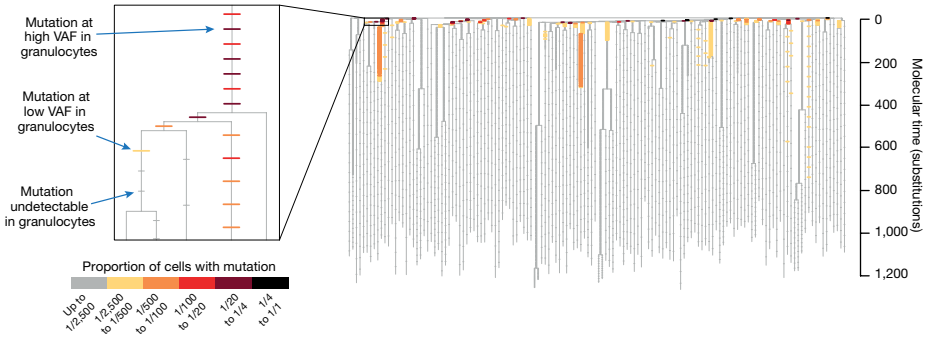


Fig. 4 | Recapture of mutations by targeted sequencing. The phylogenetic tree of cells is shown as in Fig. 2, but information from targeted sequencing of peripheral blood granulocytes from the nine-month time point is overlaid. This is shown more clearly in the inset, which zooms in on a portion of the tree. The underlying structure of the tree is shown in grey. Horizontal bars indicate the locations of every mutation in the bait set for targeted sequencing. Bars are coloured according to the proportion of cells in the sample that carry the mutations (obtained by multiplying the variant allele fraction (VAF) for autosomal mutations by two). Undetectable mutations are coloured grey and are shown as smaller bars. Mutations have been spaced evenly along a branch according to their mean variant allele fraction from the combined targeted sequencing of all granulocyte and lymphocyte time points. A higher density of baits were designed for branches that were shared by more than one colony. On these branches, the mutations are so close together that they appear as one continuous bar.

typically approximately 1/2,000. Second, most branches at the top of the phylogenetic tree have mutations that can be detected in granulocytes. This indicates that many stem cells that are not closely related to each other actively contribute to haematopoiesis during the investigated time period. If only one or a few stem cells produce granulocytes, which are replaced over time as they become exhausted, we would expect only the branches of the currently active stem cells to be detectable in granulocytes. Third, some branches contribute more to haematopoiesis than most, as indicated by a handful of branches with mutations that were detectable much further down the tree than average (although still representing a relatively small proportion: <0.6% of granulocytes). Assuming that no undiscovered driver mutations are present, this would be a consequence of genetic drift.

We developed an approximate Bayesian computation framework to quantify key properties of the stem cell compartment (Supplementary Methods and Supplementary Information). In brief, we generated 200,000 simulations of neutral haematopoiesis in adulthood (Fig. 5), varying the number of stem cells and the time between successive stem cell renewal divisions. The following assumptions were used: all active stem cells have an equal probability of dividing per unit time; the size of this pool is constant over adulthood; all active stem cells produce similar numbers of granulocytes; and stem cells accumulate somatic mutations stochastically. We then recapitulated our experimental design on each simulation in silico. Using informative summary statistics, we compared the properties of the simulations to the observed data, looking for the combinations of stem cell numbers and symmetrical cell division rate that could most closely replicate the observed data.

Under our model, the 90% credibility interval for the number of stem cells actively contributing to circulating granulocytes at one time was 44,000–215,000 (Fig. 5 and Extended Data Fig. 7). Furthermore, the estimated time between successive self-renewal stem cell divisions is most credibly in the range of 2–20 months. Simulations within this range (Extended Data Fig. 7n) most closely resembled our observed data (Extended Data Fig. 7m), whereas simulations from outside the plausible ridge showed marked differences (Extended Data Fig. 7o, p). We note the uncertainties in these estimates, which could be improved by investigating more subjects; more stratified sampling (cells from spleen, thymus and other regions of the bone marrow); sequencing more colonies in the capture phase (we now know that our 140 colonies represent only one-thousandth of the active stem cell pool); and increasing the sensitivity of mutation detection in the recapture phase.

Clonal contributions to granulocytes and lymphocytes

One of the defining features of a stem cell is its contribution to multiple cell types. Alongside granulocytes, we deep-sequenced samples of peripheral blood B and T lymphocytes (Fig. 6 and Extended

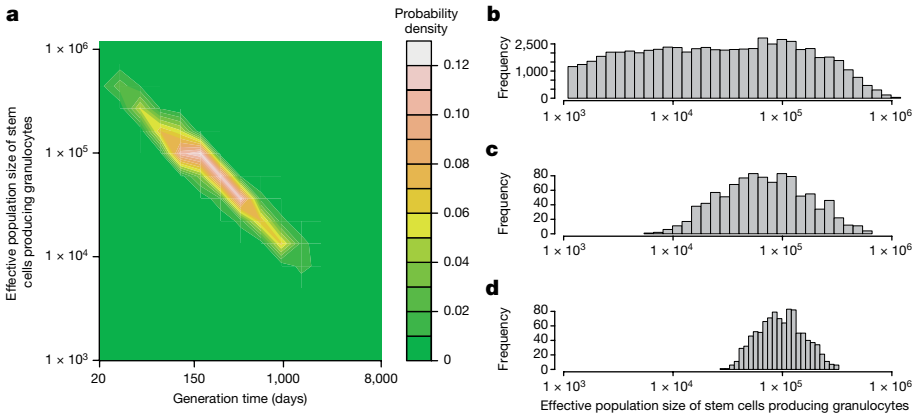


Fig. 5 | Approximate Bayesian computation of the number of stem cells and their replication rate. a, A contour plot of the most likely values for stem cell numbers and time between symmetrical stem cell divisions over the sample space that was simulated. It shows the stem cell number and generation times of the 500 simulations that produced summary statistics that were most similar to the summary statistics extracted from the observed data. **b**, The prior distribution for the number of stem cells contributing to granulocytes for the approximate Bayesian computation. 90% credibility interval: 1,696–368,457. **c**, The distribution of stem cell numbers for the 1,000 simulations that produced summary statistics most similar to the observed summary statistics. 90% credibility interval: 15,995–295,189. **d**, The posterior distribution of a neural network regression run on these 1,000 simulations. 90% credibility interval: 44,511–215,417.

© 2018 Springer Nature Limited. All rights reserved.

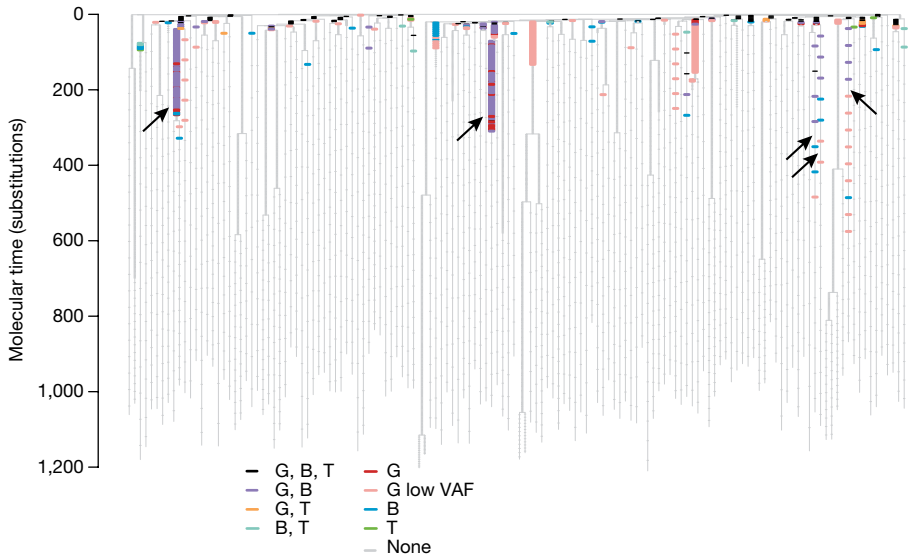


Fig. 6 | Targeted sequencing of granulocyte and lymphocyte samples. The phylogeny is depicted as in Fig. 4, with the underlying structure of the tree shown in grey, and horizontal bars represent every mutation in the bait set. Here the colouring of mutations reflects in which peripheral blood cell fractions they could be detected, as indicated by the colour key. Two colours are used for granulocytes: red for mutations only detected in granulocytes that had a sufficiently high allele fraction to be found in the shallower lymphocyte sequencing dataset and pink for mutations that were only detected in granulocytes, but at such a low allele fraction (<1 in 2,000 reads) that if they had been present in lymphocytes at this allele fraction they would not have been detected. Arrows indicate adult clones with multilineage output. B, B lymphocytes; G, granulocytes; T, T lymphocytes; VAF, variant allele fraction.

Data Fig. 9). Most of the early mutations, at the top of the phylogenetic tree, were detectable in all cell types that were investigated (black branches in Fig. 6), suggesting that there is a shared common ancestry of lymphocytes and granulocytes during development. Beyond 100 mutations in molecular time, when population size reached a plateau (Fig. 3), 464 mutations distributed across 39 branches could be detected, of which 217 on 12 branches were detected in more than one cell type.

Some adult stem cell clones contributed to detectable numbers of granulocytes and B lymphocytes, whereas their mutations were not detectable in T lymphocytes (Extended Data Fig. 9b–f). We had equivalent sequencing coverage in B and T lymphocytes, so the discrepancy is not technical. This finding suggests that descendants of these particular stem cell lines contribute a higher fraction of currently produced granulocytes and B cells than T cells. Owing to the low fraction of stem cells that were sequenced in the capture phase, we cannot exclude the existence of stem cell lines that currently contribute to all three cell types, as found in age-related clonal haematopoiesis[35].

Conclusions

The mosaic contribution of many embryonic clones to haematopoiesis and the large number of clones active in adulthood agrees well with lineage-marking studies in mice[8,9,32]. We estimate that the number of stem cells contributing to unperturbed human haematopoiesis at any given time is most probably in the hundreds of thousands and that the time between symmetric stem cell divisions falls in the range of 2–20 months; the latter estimate is similar to previous inferences[18].

We demonstrate the existence of adult human stem cell clones with multilineage output in vivo. Branches with mutations detectable in granulocytes and B lymphocytes exist beyond 300 mutations of molecular time, which would represent our subject's late teens or early twenties. The similar allele fractions in B cells and granulocytes suggest ongoing contribution of these stem cell lines to the two cell types. The key principle here is that under neutral dynamics, the drift of these branches to an appreciable proportion of granulopoiesis would have been gradual. If myeloid and B lymphocyte lineages separated early in life and underwent genetic drift independently, then we would not expect the exact same lines-of-descent to become enriched in both populations. Instead, the parsimonious explanation is that the pool of haematopoietic stem cells contributes to myeloid and B lymphocyte populations throughout life, such that the same genetic drift emerges in both lineages.

By contrast, we do not observe shared mutations between granulocytes and T lymphocytes much beyond 100 mutations of molecular time. Several possible explanations for this include a deeper separation between myeloid and T cell production[36] than between myeloid and B cells; a large and long-lived pool of T lymphocytes that dilute any ongoing contribution to T lymphopoiesis from recent stem cells; or a lower number of stem cells actively contributing to T lymphopoiesis, requiring more than the 140 colonies that were sequenced here to uncover their contribution.

The 40 trillion cells in the human body all trace their ancestry back through a series of cell divisions to the fertilised egg. All of these cells can be visualized on a single phylogenetic tree with the fertilized egg at its root. Establishing the exact position of any given cell on this tree requires a unique and permanent mark stamped on each cell with each division. Somatic mutations provide such a mark. Layering phenotypic information—be it transcriptional state, lineage output or functional resilience—on the phylogeny will enable estimation of the heritability from mother cell to daughter cell of germane somatic phenotypes. Humans can be directly researched, as the ubiquity and permanence of somatic mutations enables studies through youth and old age, in steady state and after perturbation, across blood and other organs, and in health and disease.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability

Custom R scripts and their input data central to this analysis are available on GitHub at https://github.com/HLee-Six/HSC_phylodynamics. All other codes are available from the corresponding authors upon reasonable request.

Data availability

Whole-genome and targeted sequencing data have been deposited in the European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/). Whole-genome sequencing data have been deposited with EGA accession

© 2018 Springer Nature Limited. All rights reserved.

number EGAD00001004086 and targeted sequencing data with accession number EGAD00001004087. Substitution calls have been deposited on Mendeley Data ('Population dynamics of human blood inferred from spontaneous somatic mutations': https://doi.org/10.17632/yzjw2stk7f.1). Simulated datasets (from the approximate Bayesian computation) are available from the corresponding authors upon reasonable request.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0497-0.

Received: 28 February 2018; Accepted: 1 August 2018;
Published online 5 September 2018.

1. Till, J. E. & McCulloch, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14**, 213–222 (1961).
2. Becker, A. J., McCulloch, E. A. & Till, J. E. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197**, 452–454 (1963).
3. Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917–927 (1986).
4. Kim, S. et al. Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. *Cell Stem Cell* **14**, 473-485 (2014).
5. Naik, S. H. et al. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229–232 (2013).
6. Koelle, S. J. et al. Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants. *Blood* **129**, 1448–1457 (2017).
7. Abkowitz, J. L., Catlin, S. N. & Guttorp, P. Evidence that hematopoiesis may be a stochastic process in vivo. *Nat. Med.* **2**, 190–197 (1996).
8. Sun, J. et al. Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
9. Busch, K. et al. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542–546 (2015).
10. Sawai, C. M. et al. Hematopoietic stem cells are the major source of multilineage hematopoiesis in adult animals. *Immunity* **45**, 597–609 (2016).
11. Bystrykh, L. V., Verovskaya, E., Zwart, E., Broekhuis, M. & de Haan, G. Counting stem cells: methodological constraints. *Nat. Methods* **9**, 567–574 (2012).
12. Fialkow, P. J., Gartler, S. M. & Yoshida, A. Clonal origin of chronic myelocytic leukemia in man. *Proc. Natl Acad. Sci. USA* **58**, 1468–1471 (1967).
13. Fialkow, P. J., Jacobson, R. J. & Papayannopoulou, T. Chronic myelocytic leukemia: clonal origin in a stem cell common to the granulocyte, erythrocyte, platelet and monocyte/macrophage. *Am. J. Med.* **63**, 125–130 (1977).
14. Cartier, N. et al. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**, 818–823 (2009).
15. Biasco, L. et al. In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* **19**, 107–119 (2016).
16. Notta, F. et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
17. Werner, B. et al. Reconstructing the in vivo dynamics of hematopoietic stem cells from telomere length distributions. *eLife* **4**, e08687 (2015).
18. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460–4466 (2011).
19. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
20. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
21. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
22. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
23. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550–555 (2018).
24. Notta, F. et al. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218–221 (2011).
25. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
26. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
27. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
28. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
29. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
30. Plusa, B. et al. The first cleavage of the mouse zygote predicts the blastocyst axis. *Nature* **434**, 391–395 (2005).
31. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
32. Pei, W. et al. *Polylox* barcoding reveals haematopoietic stem cell fates realized *in vivo*. *Nature* **548**, 456–460 (2017).
33. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
34. Lan, S., Palacios, J. A., Karcher, M., Minin, V. N. & Shahbaba, B. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* **31**, 3282–3289 (2015).
35. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).
36. Schlenner, S. M. et al. Fate mapping reveals separate origins of T cells and myeloid lineages in the thymus. *Immunity* **32**, 426–436 (2010).

Acknowledgements This work was supported by the Leukemia Lymphoma Society, the WBH Foundation and the Wellcome Trust. P.J.C. is a Wellcome Trust Senior Clinical Fellow (WT088340MA). H.L.-S. is a recipient of a Wellcome Trust PhD studentship. N.F.Ø. is the recipient of a Danish Lundbeck Fellowship (2016-17) and M.S.S. is the recipient of a BBSRC CASE Industrial PhD Studentship. Work in the D.G.K. laboratory is supported by a Bloodwise Bennett Fellowship (15008), a European Research Council Starting Grant (ERC-2016-STG–715371) and a European Hematology Association Non-Clinical Advanced Research Fellowship. Work in the A.R.G. laboratory is supported by the Wellcome Trust, Bloodwise, Cancer Research UK, the Kay Kendall Leukaemia Fund, and the Leukemia and Lymphoma Society of America. Work in E.L. laboratory is supported by a Wellcome Trust Sir Henry Dale Fellowship, BBSRC and a European Haematology Association Non-Clinical Advanced Research Fellowship. The D.G.K., E.L. and A.R.G. laboratories are supported by a core support grant from the Wellcome Trust and Medical Research Council to the Cambridge Stem Cell Institute. We acknowledge further assistance from the National Institute for Health Research Cambridge Biomedical Research Centre and the Cambridge Experimental Cancer Medicine Centre. We thank R. Grenfell and M. Strzelecki in the Flow Cytometry Facility of the Cancer Research UK Cambridge Institute for technical assistance and suggestions; C. Eaves and P. Beer for discussions about the experimental design; P. Goodell for discussions about the results; J. Grinfeld and C. Gonzalez-Arias for drawing peripheral blood; and G. Collord for providing negative control samples for targeted sequencing.

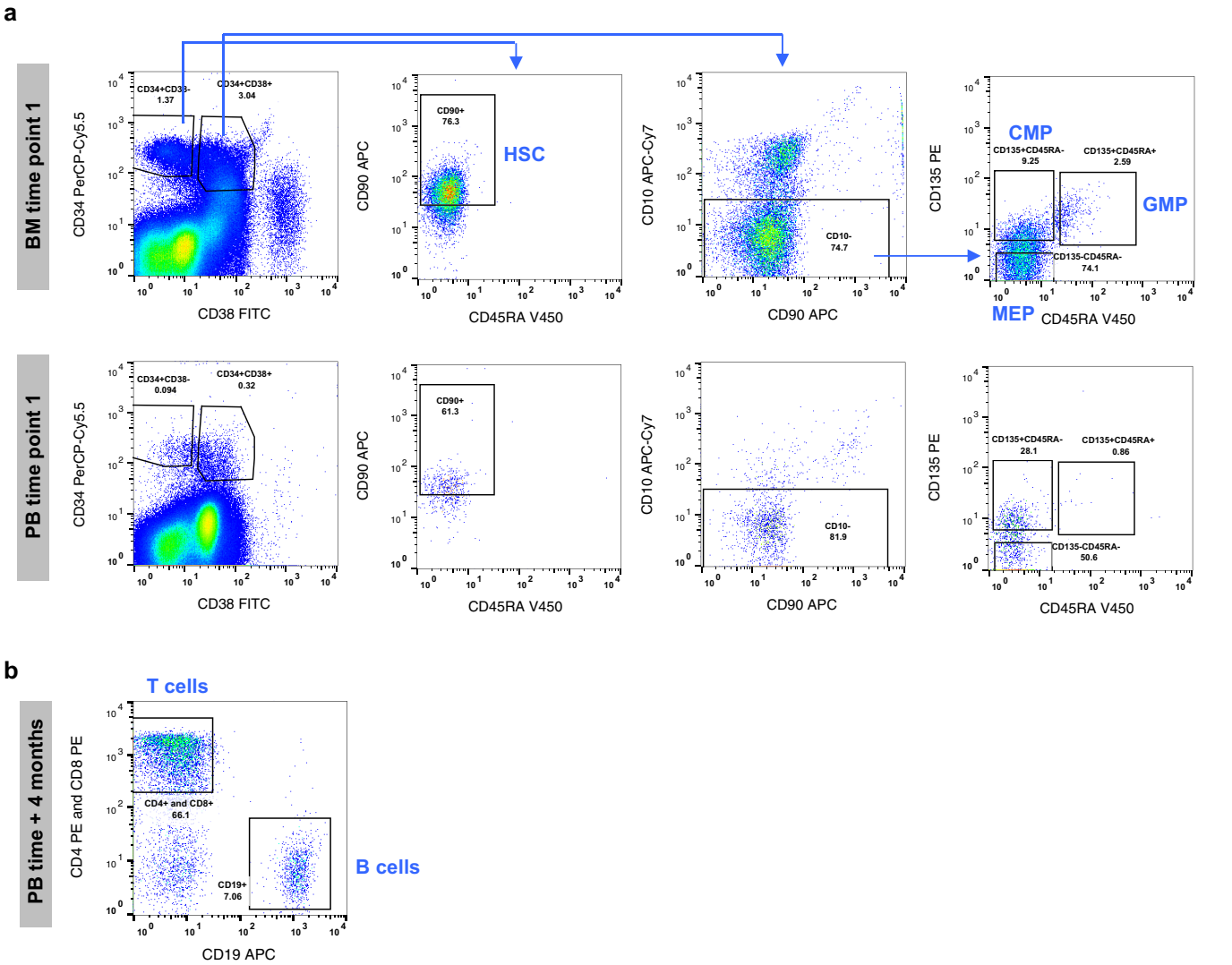
Reviewer information *Nature* thanks H.-R. Rodewald, L. Shlush and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.J.C., D.G.K., A.R.G. and H.L.-S. designed the experiments. B.J.P.H. performed the bone marrow aspiration experiments. D.G.K., N.F.Ø., M.S.S. and M.B. cultured HSPCs, with advice and reagents from E.L., and isolated cells from subsequent blood draws. E.A. and L.O. oversaw DNA extraction and library preparation. H.L.-S. performed most of the data curation and statistical analysis. S.G. helped to call mutations and construct the phylogeny. I.M. tested for positive selection. M.R.S. oversaw the analysis of mutational signatures. R.J.O. and P.J.C. separated signal from noise in targeted sequencing analysis. K.D. helped to design the approximate Bayesian computation along with P.J.C. and H.L.-S. and co-wrote the supplementary results on inference of population size. H.L.-S., P.J.C., D.G.K. and A.R.G. wrote the manuscript, with contributions from all authors.

Competing interests The authors declare no competing interests.

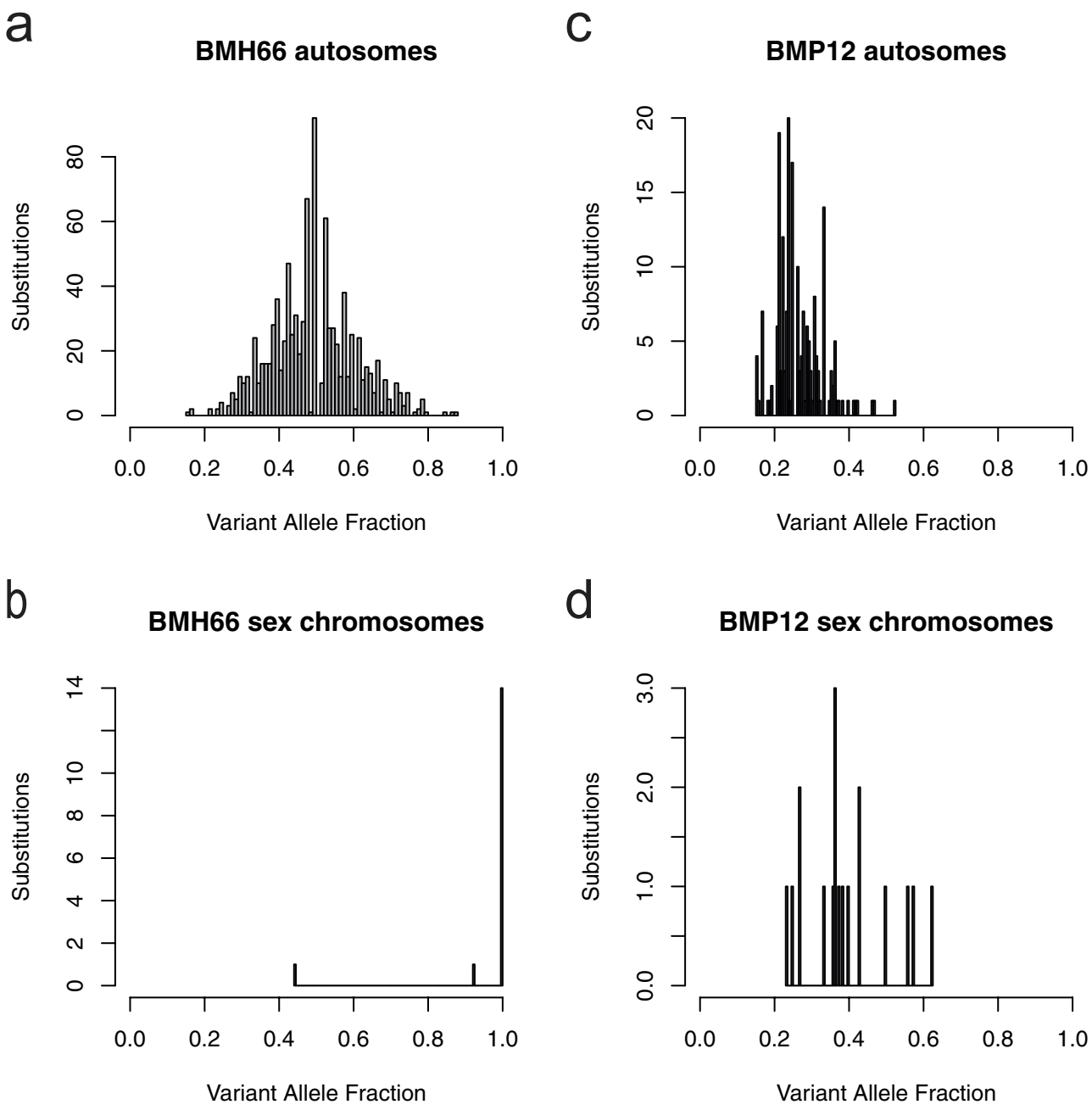
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0497-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0497-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.R.G., D.G.K. and P.J.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.



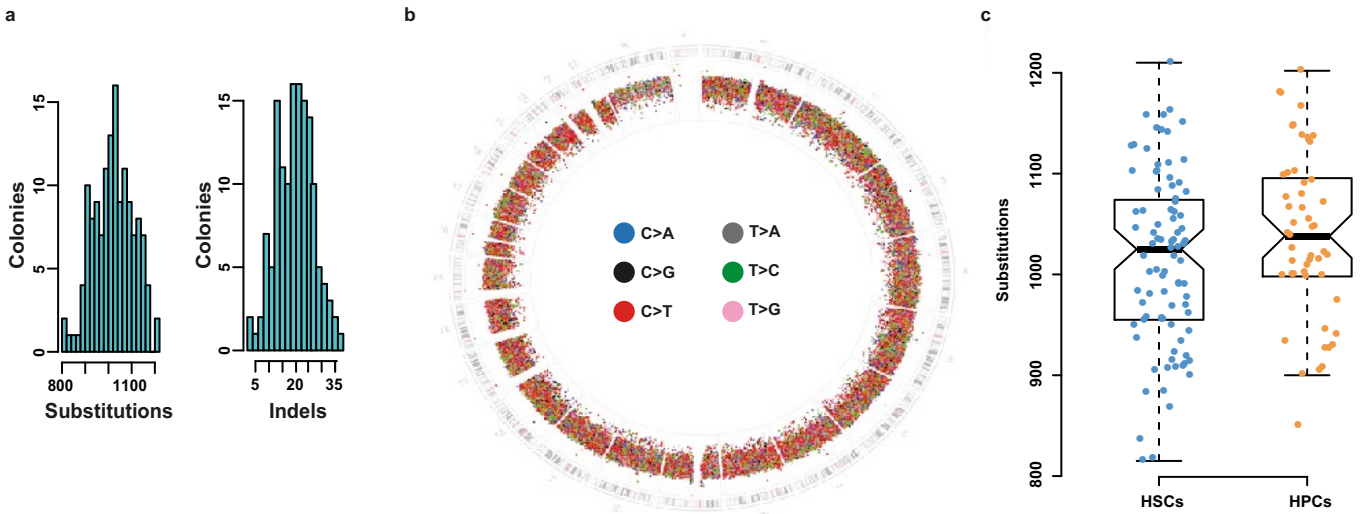
Extended Data Fig. 1 | Cell sorting strategy. a, Sorting of stem and progenitor cells. Human bone marrow (BM) and peripheral blood (PB) mononuclear cells (time point 1) were stained with anti-CD34, anti-CD38, anti-CD45RA, anti-CD90, anti-CD10 and anti-CD135 antibodies. After exclusion of debris and doublets, gatings on CD34, CD38 and CD90 were used to separate $CD34^{+}CD38^{-}CD90^{+}CD45RA^{-}$ HSCs. The $CD34^{+}CD38^{+}$ compartment was gated for $CD10^{-}$ cells before gating on CD135 (also known as FLT3) and CD45RA to separate progenitor compartments: $CD135^{+}CD45RA^{-}$ common myeloid progenitor (CMPs), $CD135^{+}CD45RA^{+}$ granulocyte–macrophage progenitors (GMPs) and $CD135^{-}CD45RA^{-}$ megakaryocyte–erythrocyte progenitors (MEPs). **b**, Sorting of B and T lymphocytes. Peripheral blood mononuclear cells (time point, after 4 months) were stained with anti-CD4, anti-CD8 and anti-CD19 antibodies. After exclusion of debris and doublets, the $CD4^{+}CD8^{+}CD19^{-}$ gate was used to isolate T cells, while the $CD4^{-}CD8^{-}CD19^{+}$ gate was used to isolate B cells. $n = 20{,}000$ events.

© 2018 Springer Nature Limited. All rights reserved.



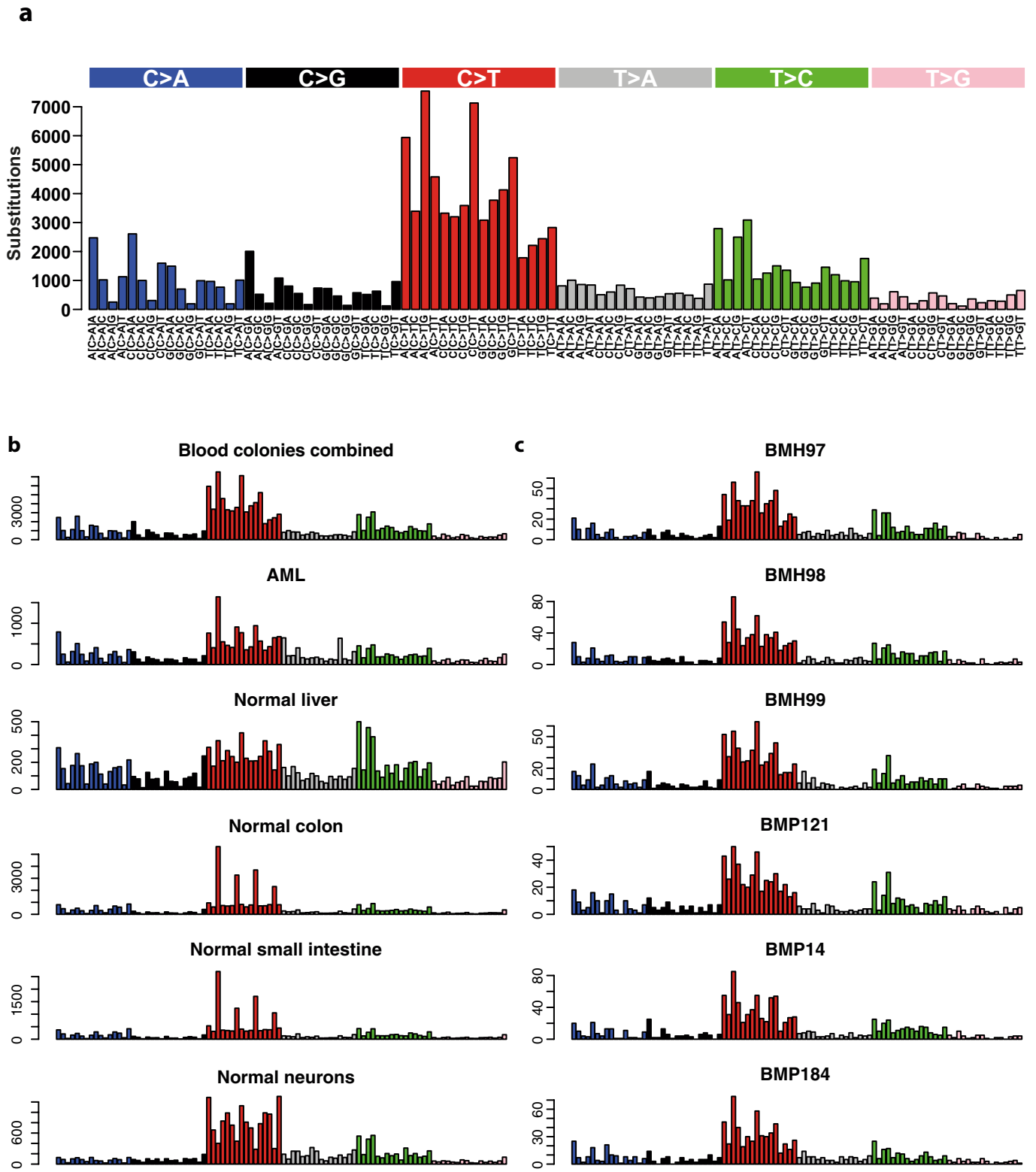
Extended Data Fig. 2 | Quality control of colonies from single-cell derived clones. Example histograms of the variant allele fraction (VAF—the proportion of sequencing reads that report the mutation) of mutations in single colonies. **a**, The VAF of all mutations on autosomes in a typical clonal colony. Because there are two copies of each autosome, and each mutation occurs on only one of them, in a clonal sample the VAF of autosomal mutations is binomially distributed with a mean of 0.5. **b**, The VAF of all mutations on the X chromosome in the same clonal colony. Because the subject is male, there is only one copy of the X chromosome, and so true mutations here must have a VAF of 1. Occasionally, lower VAFs are seen when a mutation is not detected on a read, when a read from another locus is aberrantly mapped to the locus in question and thus lowering the apparent coverage, or when a mutation is acquired in vitro. **c**, **d**, The VAF of autosomal and X chromosome mutations, respectively, in a typical colony seeded by more than one cell. As not all the reads come from the same cell, and most mutations are private to a given cell, a lower proportion of DNA molecules carry the mutation in a polyclonal colony than in a clonal colony, resulting in a leftward shift of the peak of the VAF histogram. These histograms suggest that the number of mutations acquired by the colonies after a few weeks of in vitro expansion is a small fraction of those acquired in vivo over 60 years of life.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Mutation burden of colonies. a, A histogram of substitution (left) and indel (right) burden per colony. **b**, The location around the genome of substitutions from all clones combined is shown as a circos plot. The outermost ring of the circos plot depicts the karyotypic ideogram. Moving inwards, base substitutions are shown as rainfall plots in which the height of the dot in the substitution ring is proportional to the $\log_{10}$ of the distance to the next mutation and with the colour of the dot illustrating the base change, as shown in the key. **c**, A comparison of the substitution burden between stem cells and progenitor cells. There were not significantly more mutations in progenitors than stem cells ($P = 0.14$, Wilcoxon rank-sum test).

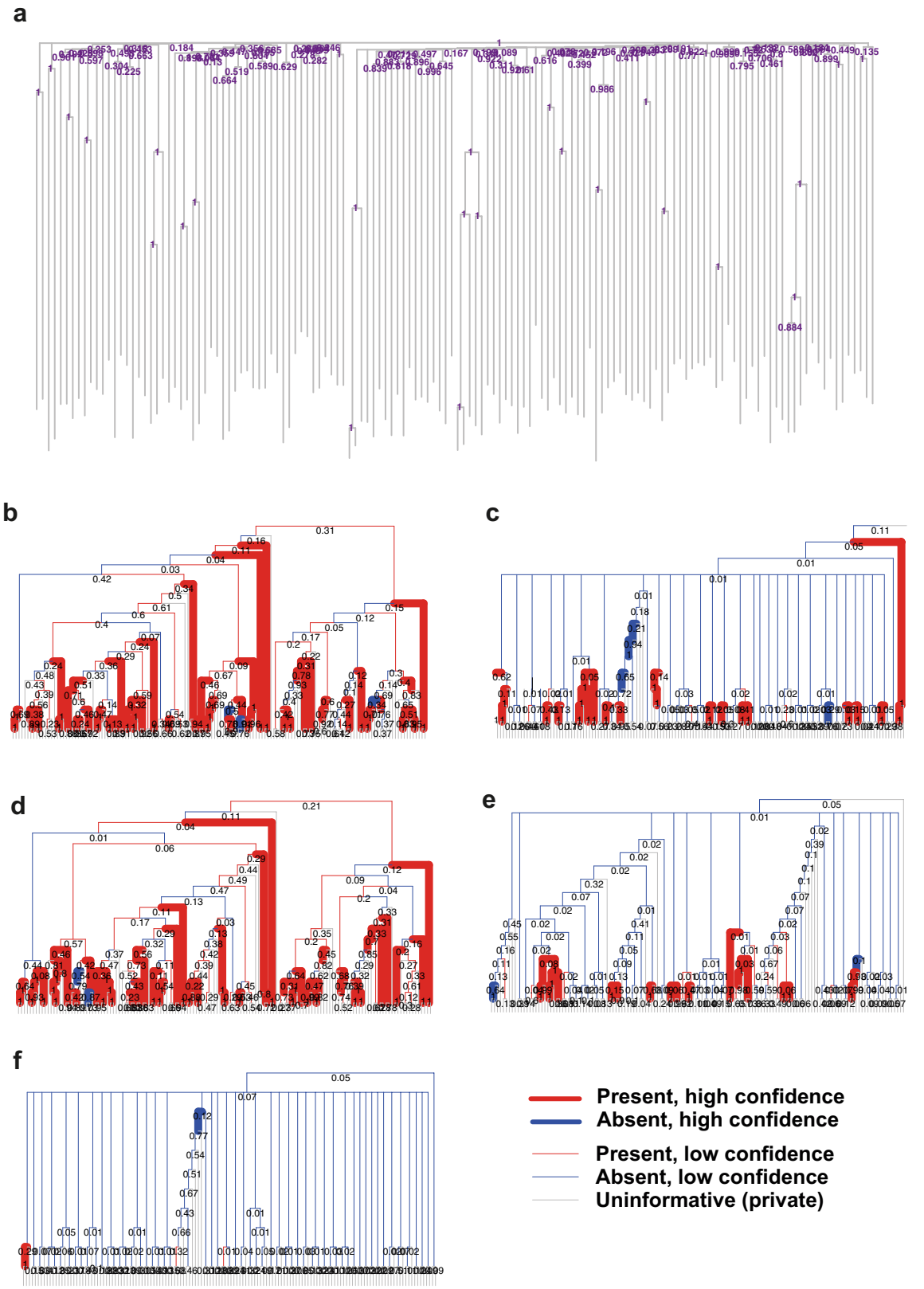
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Trinucleotide context of mutations in normal blood colonies. a, The trinucleotide context of substitutions for all colonies combined. Substitutions can be classed according to the base change (referred to by the pyrimidine of the mutated base pair), and the bases 5′ and 3′ of the mutated one, into 96 categories. The counts in each of these categories are shown. **b**, Comparison with pooled acute myeloid leukaemia genomes, excluding genomes with >1,500 mutations, and publicly available data on normal tissues that have been whole-genome sequenced so far. The ordering of bars is the same as in **a**, and the same figure as in **a** is provided again at the same resolution of the other data for ease of comparison. Please note that these samples have been sequenced on different platforms using different systems, which is likely to result in small differences. Normal liver, normal colon and normal small intestine data were obtained from whole-genome sequencing of single-cell-derived organoids[19], whereas normal neurons were derived from single cells that had undergone whole-genome amplification[37]. **c**, Example trinucleotide substitution plots for a selection of individual colonies derived from either stem cells (which have the prefix BMH) or progenitor cells (which have the prefix BMP). The ordering of bars is the same as in **a**.

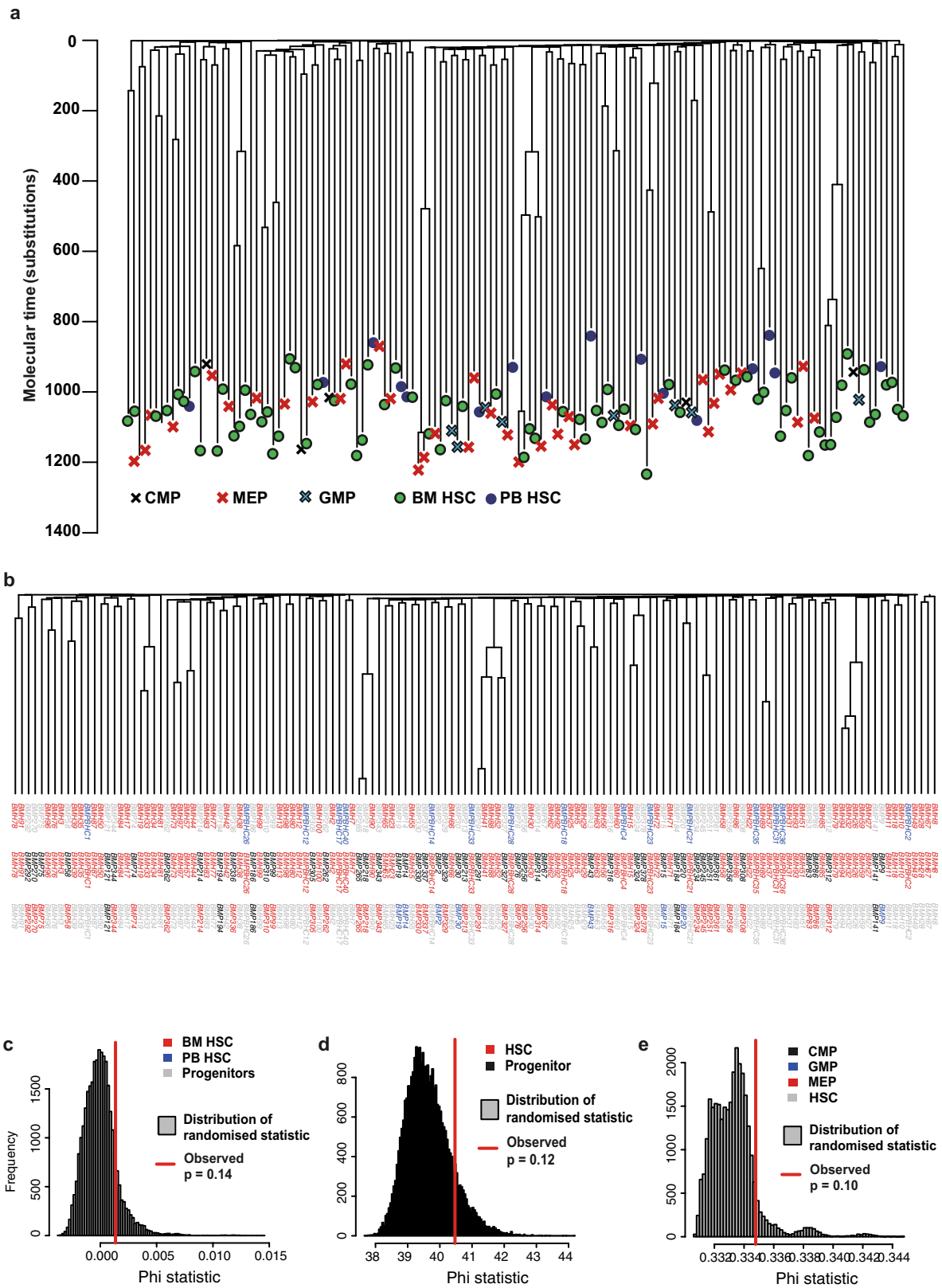
37. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).

© 2018 Springer Nature Limited. All rights reserved.



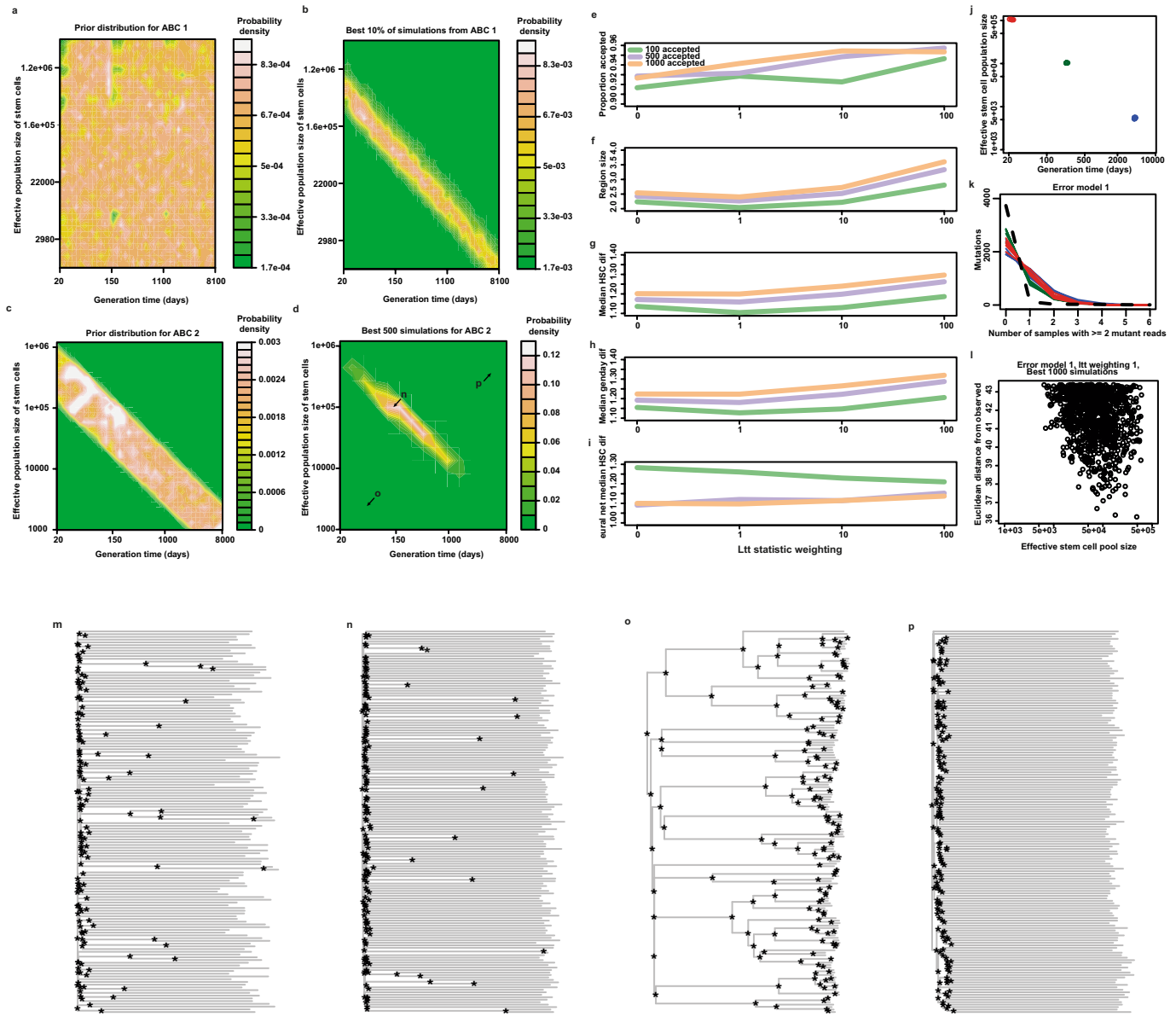
Extended Data Fig. 5 | Construction of the phylogeny using different methods. a, The phylogeny of cells as presented in Figs. 2, 4, 6, but with the addition of *P* values next to every node, derived by bootstrapping the substitution matrix 1,000 times, building a tree using SCITE for each replicate, and counting the proportion of the bootstrapped trees that support each node. **b**–**f**, Phylogenies constructed using different datasets and methods. In each case the phylogeny was constructed using 100 bootstraps of the data, and the *P* value for each node shown underneath it. Branches are coloured by whether a branch ancestral to exactly the same descendants is also present in the SCITE tree, and are drawn with a thicker line if the branch is recovered in ≥70% of bootstrap replicates. **b**, Substitution and indel datasets combined, building the tree by maximum parsimony. **c**, Substitution, indel and neighbour-joining datasets combined, building the tree by neighbour joining. **d**, Substitutions, tree build by maximum parsimony. **e**, Indels, tree built by maximum parsimony. **f**, Short tandem repeats, tree built by neighbour joining.

© 2018 Springer Nature Limited. All rights reserved.



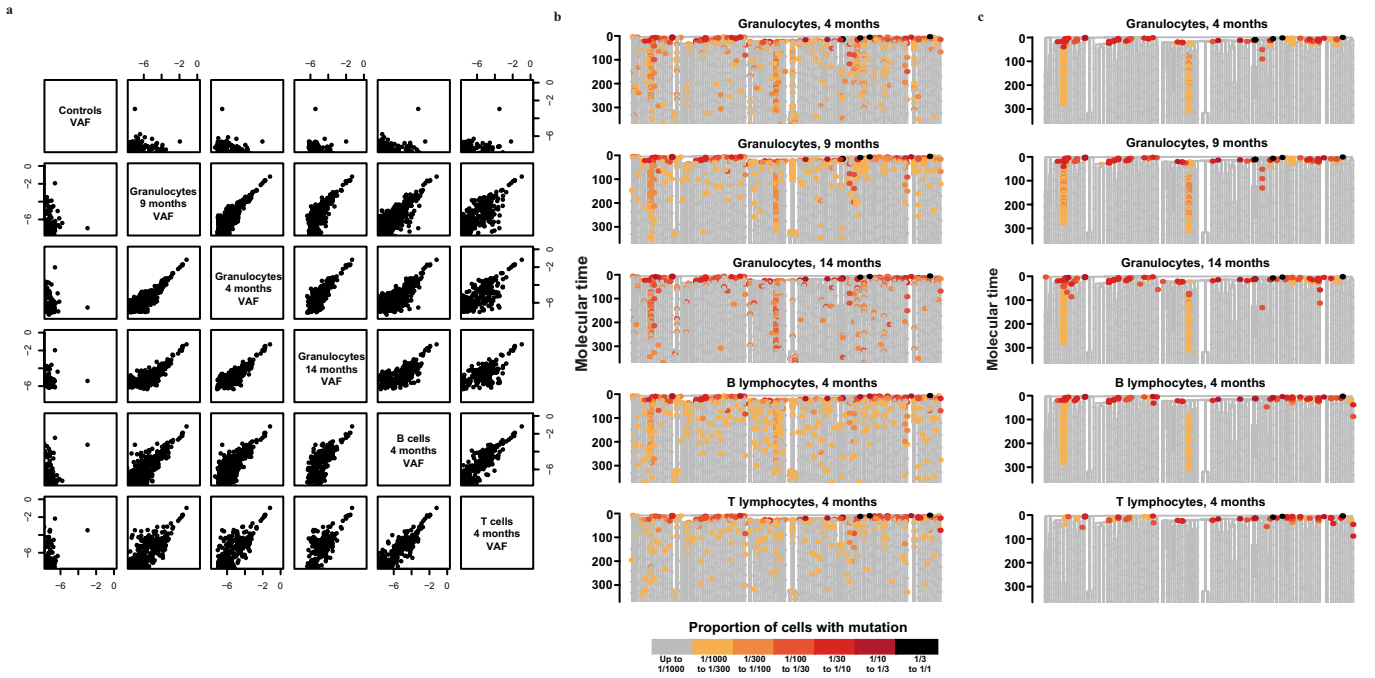
Extended Data Fig. 6 | Relationship between cell types in the phylogeny. **a**, The phylogeny showing different stem and progenitor cell types. **b**, The phylogeny is shown as in **a**, but with the labels underneath coloured according to which cell types are being compared. The first row of labels has stem cells from bone marrow in red, progenitor cells from bone marrow in grey and stem cells from peripheral blood in black. The second row of labels has stem cells in red and bone marrow progenitors in black. The third row of labels has MEPs in red, CMPs in black, GMPs in blue and stem cells in grey. **c**–**e**, Analysis of molecular variance is used to test for clustering on the phylogeny for stem cells derived from peripheral blood versus bone marrow cells (**c**), stem cells versus progenitors (**d**) and different progenitor types (**e**). In each panel, a histogram of the null distribution of the statistic used to detect clustering is shown. Distributions were obtained by randomly permuting which cells were assigned to which category. Comparisons are only between cell types not shown in grey in **b**. The observed value of the statistic is shown as a red vertical line.

© 2018 Springer Nature Limited. All rights reserved.



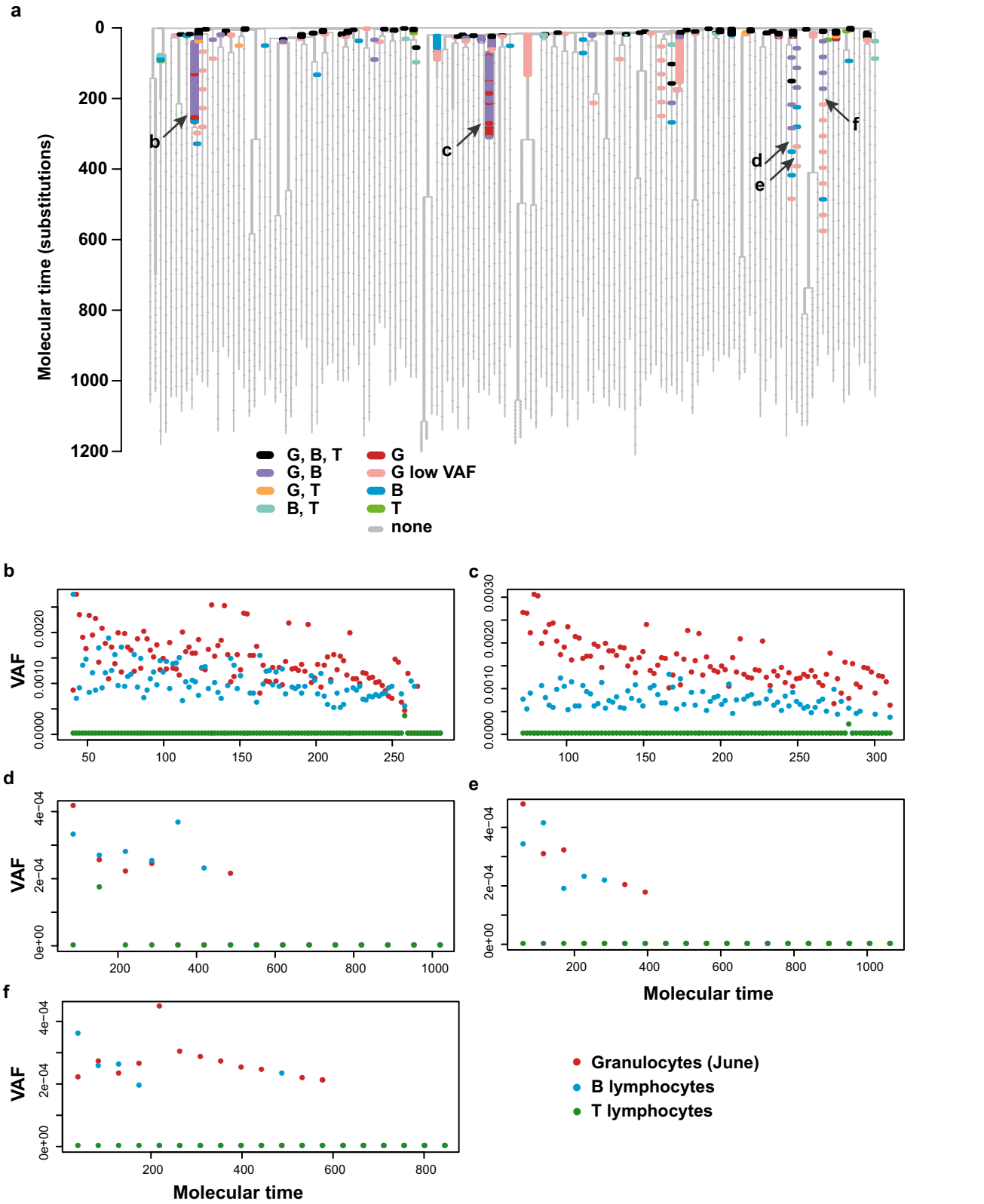
Extended Data Fig. 7 | Approximate Bayesian computations. a, The joint prior distribution for stem cell numbers (HSCs) and the generation time for the first approximate Bayesian computation (ABC). **b**, The location in sample space of the 10% of simulations that produced summary statistics (using only the ltt summary statistics; Supplementary Methods and Supplementary Information) most similar to the observed summary statistics. **c**, The joint prior distribution for the second ABC, in the area of sample space indicated to be plausible by the first set of simulations. **d**, The joint posterior distribution of the best 500 simulations from the second ABC, as shown in Fig. 5 for ease of reference. 'n', 'o' and 'p' on the plot indicate the position in sample space from which panels **n**–**p** were drawn. **e**–**i**, Cross-validation of the model to choose the number of accepted simulations and the weighting applied to the ltt summary statistics (Supplementary Methods and Supplementary Information). **j**, For illustrative purposes, five simulations were sampled for each of three population sizes along the plausible diagonal of sample space indicated in **b**. One set of summary statistics are shown for these simulations in **k**. **k**, The red line indicates a simulation coming from the area of sample space indicated by a red point in **j**; and similarly for blue and green lines. The black dotted line indicates the observed values for these summary statistics. These summary statistics provide a count—for the different numbers of samples (*x* axis)—of how many of the 3,952 mutations that we considered (*y* axis) are in this many samples with two or more reads, using error model 1 (which simulates errors according to the error rate in control DNA (Supplementary Methods)). The same summary statistics were calculated for different mutant read number cut-offs. **l**, For each of the 1,000 simulations that produce summary statistics that were the most similar to the observed data, the Euclidean distance from the observed data (*y* axis) is plotted against the number of stem cells in that simulation (*x* axis). This information is used by the neural network regression step to define the most likely value for the number of stem cells. The most similar values are seen at around 100,000 stem cells, which was the location of the median of the posterior distribution from neural network regression. **m**, The observed phylogeny, with branch points indicated by asterisks. **n**–**p**, Phylogenies drawn from simulations that occur at the points in sample space indicated in **d**. **n**, A relatively plausible simulation, since the pattern of branch points is not dissimilar from the pattern of the observed phylogeny (**m**). Simulations with smaller stem cell populations and faster stem cell turnover rates resulted in phylogenies in which the stem cells were very closely related to each other (**o**), whereas those with larger populations and slower turnover result in phylogenies in which the stem cells only share an embryonic common ancestor, and no branches are seen through the tree (**p**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Targeted sequencing data. a, Correlations between the VAFs of all sequenced samples, shown on a log scale. Note that samples that were sequenced to a lower depth cannot have VAFs as small as samples sequenced to higher depths. **b**, Targeted sequencing information with no error correction. The data are shown as in Fig. 4 for all analysed samples, but focusing on only the first 350 mutations of molecular time. To allow a better comparison between samples that were sequenced at different depths, a higher detection threshold and different detection threshold are used relative to Fig. 4. **c**, Targeted sequencing information after using cord blood controls for sequencing error correction with the Bayesian generalized Poisson mixed-effects model. The colour scale is the same as in **b**. The data for the granulocytes at the nine-month time point are the same as in Fig. 4 (provided again for ease of comparison), but plotted with a different colour scale.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Multilineage clonal output. a, The phylogeny with targeted sequencing information in different blood fractions overlaid as in Fig. 6, shown again here for ease of reference. The colouring of mutations reflects in which peripheral blood cell fractions they could be detected, as indicated by the colour key. Arrows indicate adult clones with multilineage output, with letters corresponding to panels **b**–**f**. B, B lymphocytes; G, granulocytes; G low VAF, granulocytes, allele fraction too low to be detected in lymphocytes; T, T lymphocytes. **b**–**f**, VAFs of all mutations on branches (indicated by arrows in **a**) with mutations beyond molecular time 100 that are detectable in granulocytes and B lymphocytes but not in T lymphocytes.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0531-2

Combination therapy with anti-HIV-1 antibodies maintains viral suppression

Pilar Mendoza[1,19], Henning Gruell[2,3,4,19], Lilian Nogueira[1], Joy A. Pai[1], Allison L. Butler[1], Katrina Millard[1], Clara Lehmann[3,4,5], Isabelle Suárez[3,4,5], Thiago Y. Oliveira[1], Julio C. C. Lorenzi[1], Yehuda Z. Cohen[1], Christoph Wyen[3,6], Tim Kümmerle[3,6], Theodora Karagounis[1], Ching-Lan Lu[1], Lisa Handl[7], Cecilia Unson-O'Brien[1], Roshni Patel[1], Carola Ruping[2], Maike Schlotz[2], Maggi Witmer-Pack[1], Irina Shimeliovich[1], Gisela Kremer[3], Eleonore Thomas[3], Kelly E. Seaton[8], Jill Horowitz[1], Anthony P. West Jr[9], Pamela J. Bjorkman[9], Georgia D. Tomaras[8,10,11,12], Roy M. Gulick[13], Nico Pfeifer[7,14,15,16], Gerd Fätkenheuer[3,4], Michael S. Seaman[17], Florian Klein[2,4,5,20]*, Marina Caskey[1,20]* & Michel C. Nussenzweig[1,18,20]*

Individuals infected with HIV-1 require lifelong antiretroviral therapy, because interruption of treatment leads to rapid rebound viraemia. Here we report on a phase 1b clinical trial in which a combination of 3BNC117 and 10-1074, two potent monoclonal anti-HIV-1 broadly neutralizing antibodies that target independent sites on the HIV-1 envelope spike, was administered during analytical treatment interruption. Participants received three infusions of 30 mg kg^{-1} of each antibody at 0, 3 and 6 weeks. Infusions of the two antibodies were generally well-tolerated. The nine enrolled individuals with antibody-sensitive latent viral reservoirs maintained suppression for between 15 and more than 30 weeks (median of 21 weeks), and none developed viruses that were resistant to both antibodies. We conclude that the combination of the anti-HIV-1 monoclonal antibodies 3BNC117 and 10-1074 can maintain long-term suppression in the absence of antiretroviral therapy in individuals with antibody-sensitive viral reservoirs.

During infection, HIV-1 is reverse transcribed and integrated as a provirus into the host genome. Although the vast majority of infected cells die by apoptosis or pyroptosis[1], a small percentage survive and harbour transcriptionally silent, integrated proviruses that comprise a reservoir that can be reactivated. Once established, the latent reservoir has an estimated half-life of 44 months, resulting in the lifelong requirement for antiretroviral therapy (ART)[2]. Passive administration of potent broadly neutralizing monoclonal anti-HIV-1 antibodies (bNAbs) represents a potential alternative to antiretroviral drugs because, in addition to neutralizing the virus, antibodies engage the host immune system and have long half-lives[3–5].

In human clinical trials, viraemic individuals who received 3BNC117 or VRC01, two related bNAbs that target the CD4 binding site on the HIV-1 envelope spike, or 10-1074, a bNAb that targets the base of the V3 loop and surrounding glycans, showed significant reductions in viremia[6–8]. Moreover, in HIV-1-infected individuals undergoing analytical treatment interruption (ATI) of antiretroviral therapy, four infusions of 3BNC117 maintained virus suppression for a median of 10 weeks compared to 2.3 weeks in historical controls[9,10]. By contrast, six infusions of VRC01 maintained suppression for 5.6 weeks[11]. The difference in activity between VRC01 and 3BNC117 in preclinical experiments[12,13] and clinical trials[6,7,9,11] is consistent with the lower relative neutralization potency of VRC01.

Across all bNAb clinical trials to date, and similar to monotherapy with antiretroviral drugs, treatment with any single bNAb was associated with the emergence of antibody-resistant viral variants[6–9,11]. Like antiretroviral drugs, combinations of bNAbs are more effective than individual antibodies in HIV-1 infected humanized mice and simian/human immunodeficiency virus (SHIV)-infected macaques[14–16]. By contrast, antibody combinations showed little if any efficacy in suppressing viraemia during ATI in humans[17,18]. However, these earlier studies were performed using bNAbs that were less potent than 3BNC117 and 10-1074. Here we investigate whether the bNAb combination of 3BNC117 and 10-1074 can maintain viral suppression during ATI in HIV-1-infected humans.

Combination bNAb infusion is well-tolerated

To evaluate the effects of the combination of 3BNC117 and 10-1074 on maintaining HIV-1 suppression during ATI, we conducted a phase 1b clinical trial (Fig. 1a). HIV-1-infected individuals on ART were pre-screened for 3BNC117 and 10-1074 sensitivity of bulk outgrowth culture-derived viruses in an in vitro neutralization assay using TZM-bl cells[19]. Consistent with previous results, 64% and 71% of the outgrowth viruses were sensitive to 3BNC117 and 10-1074, respectively, and 48% were sensitive to both[8,9,20] (half-maximum inhibitory concentration (IC_{50}) $\leq 2\,\mu g\,ml^{-1}$; Extended Data Fig. 1a and Supplementary Table 1).

Study eligibility criteria included ongoing ART for at least 24 months with plasma HIV-1 RNA levels of <50 copies per ml for at least 18 months (one blip <500 copies per ml was allowed) and <20 copies per ml at screening, as well as $CD4^+$ T cell counts >500 cells per μl (Extended Data Figs. 1b, 2a). Enrolled participants received three infusions of 30 mg kg^{-1} of 3BNC117 and 10-1074 each at

[1]Laboratory of Molecular Immunology, The Rockefeller University, New York, NY, USA. [2]Laboratory of Experimental Immunology, Institute of Virology, University Hospital Cologne, Cologne, Germany. [3]Department I of Internal Medicine, University Hospital Cologne, Cologne, Germany. [4]German Center for Infection Research, partner site Bonn–Cologne, Cologne, Germany. [5]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. [6]Praxis am Ebertplatz, Cologne, Germany. [7]Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Tübingen, Germany. [8]Duke Human Vaccine Institute, Duke University, Durham, NC, USA. [9]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [10]Department of Surgery, Duke University, Durham, NC, USA. [11]Department of Immunology, Duke University, Durham, NC, USA. [12]Department of Molecular Genetics and Microbiology, Duke University, Durham, NC, USA. [13]Division of Infectious Diseases, Weill Cornell Medicine, New York, NY, USA. [14]Medical Faculty, University of Tübingen, Tübingen, Germany. [15]German Center for Infection Research, partner site Tübingen, Tübingen, Germany. [16]Max Planck Institute for Informatics, Saarbrücken, Germany. [17]Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [18]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [19]These authors contributed equally: Pilar Mendoza, Henning Gruell. [20]These authors jointly supervised this work: Florian Klein, Marina Caskey, Michel C. Nussenzweig. *e-mail: florian.klein@uk-koeln.de; mcaskey@rockefeller.edu; nussen@rockefeller.edu

© 2018 Springer Nature Limited. All rights reserved.

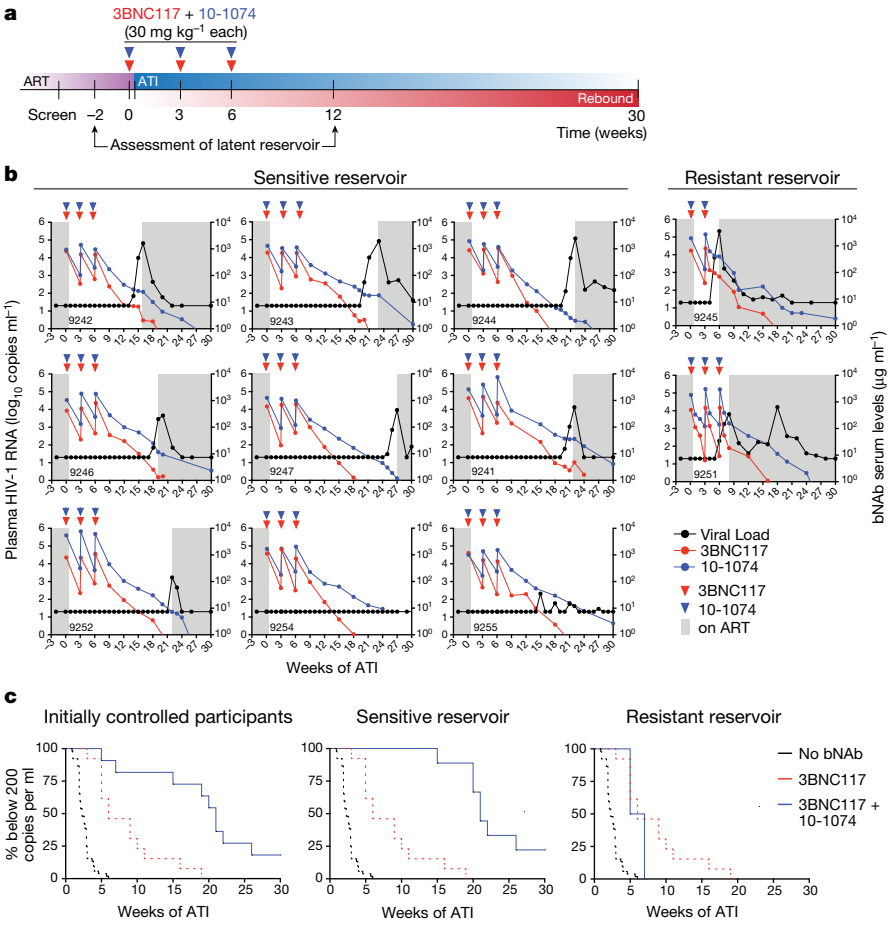


Fig. 1 | Delayed viral rebound with 3BNC117 and 10-1074 combination therapy during ATI. a, Study design. Red and blue triangles represent 3BNC117 and 10-1074 infusions, respectively. **b**, Plasma HIV-1 RNA levels (black; left *y* axis) and bNAb serum concentrations (3BNC117, red; 10-1074, blue; right *y* axis) in the nine bNAb-sensitive participants (left) and the two participants with pre-existing resistance against one of the antibodies (right). Red and blue triangles indicate 3BNC117 and 10-1074 infusions, respectively. Serum antibody concentrations were determined by TZM-bl assay. Grey-shaded areas indicate time on ART. Lower limit of detection of HIV-1 RNA was 20 copies per ml. **c**, Kaplan–Meier plots summarizing time to viral rebound for the participants with HIV-1 RNA <20 copies per ml two weeks before and at the start of ATI ($n = 11$, left), for the participants sensitive to both antibodies ($n = 9$, centre), and for the participants that showed pre-existing resistance to one of the antibodies ($n = 2$, right). The *y* axis shows the percentage of participants that maintain viral suppression. The *x* axis shows the time in weeks after start of ATI. Participants receiving the combination of 3BNC117 and 10-1074 are indicated by the blue line. Dotted red lines indicate a cohort of individuals receiving 3BNC117 alone during ATI[9] ($n = 13$) and dotted black lines indicate a cohort of participants who underwent ATI without intervention[10] ($n = 52$).

three-week intervals beginning two days before treatment interruption (Fig. 1a). Individuals whose regimens contained non-nucleoside reverse transcriptase inhibitors were switched to an integrase inhibitor-based regimen four weeks before discontinuing ART (Extended Data Figs. 1b, 2a). Viral load and $CD4^+$ T cell counts were monitored every 1–2 weeks (Supplementary Table 2). ART was reinitiated and antibody infusions were discontinued if viraemia of >200 copies per ml was confirmed. Time of viral rebound was defined as the first of two consecutive viral loads of >200 copies per ml. Fifteen individuals were enrolled, but four of them showed viral loads of >20 copies per ml two weeks before or at the time of the first bNAb infusion and they were excluded from efficacy analyses (Extended Data Fig. 1b and Supplementary Table 2).

Antibody infusions were generally safe and well-tolerated with no reported serious adverse events or antibody-related adverse events, except for mild fatigue in two participants (Supplementary Table 3). The mean $CD4^+$ T cell count was 685 and 559 cells per μl at the time of first antibody infusion and at rebound, respectively (Extended Data Fig. 2b and Supplementary Table 2). Reinitiation of ART after viral rebound resulted in resuppression of viraemia (Supplementary Table 2). We conclude that combination therapy with 3BNC117 and 10-1074 is generally safe and well-tolerated.

The serum half-life of each antibody was measured independently by TZM-bl assay and anti-idiotype enzyme-linked immunosorbent assay (ELISA, Extended Data Fig. 2c, d and Supplementary Table 2). 3BNC117 had a half-life of 12.5 and 17.6 days as measured using TZM-bl and ELISA, respectively (Extended Data Fig. 2c, d). The half-life of 10-1074 was 19.1 and 23.2 days as measured by TZM-bl and ELISA, respectively; significantly longer than 3BNC117 in both assays ($P = 0.0002$ and $P = 0.02$, Extended Data Fig. 2e, f). These measurements are similar to those observed when each antibody was administered alone in ART-treated HIV-1-infected individuals[6,8,9]. We conclude that the pharmacokinetic profiles of 3BNC117 and 10-1074 are not altered when they are used in combination.

The combination of bNAbs maintains viral suppression

For the 11 individuals who had complete viral suppression (HIV-1 RNA <20 copies per ml) during the screening period and at day 0, combination antibody therapy was associated with maintenance of viral suppression for between 5 and more than 30 weeks (Fig. 1b, c and Supplementary Table 2). The median time to rebound was 21 weeks compared to 2.3 weeks for historical controls who participated in previous non-interventional ATI studies[10] and 6–10 weeks for monotherapy with 3BNC117[9] (Fig. 1c). Together, 9 of the 11 participants maintained viral

© 2018 Springer Nature Limited. All rights reserved.

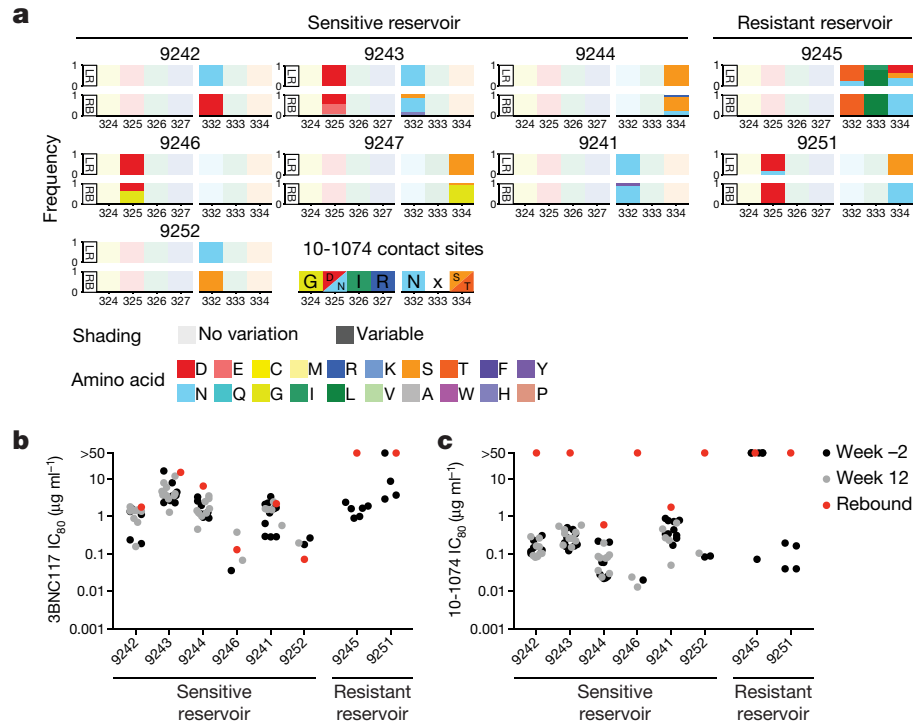


Fig. 2 | Amino acid variants at 10-1074 contact sites and bNAb sensitivity of reactivated latent and rebound viruses. a, Colour charts show Env contact sites of 10-1074 at the G(D/N)IR motif (positions 324–327, according to HXB2 numbering) and the glycan at the potential *N*-linked glycosylation site at position 332 (NxS/T motif at positions 332–334). Diagram shows the seven bNAb-sensitive participants that rebounded before week 30 (left) and the two individuals with pre-existing resistance to one of the two antibodies (right). LR, latent reservoir viruses isolated by Q^2VOA; RB, rebound viruses isolated by SGA (plasma) or viral outgrowth (PBMCs). Each amino acid is represented by a colour and the frequency of each amino acid is indicated by the height of the rectangle. Shaded rectangles indicate the lack of variation between latent reservoir virus and rebound virus at the indicated position. Full-colour rectangles represent amino acid residues with changes in distribution between reservoir and rebound viruses. **b**, **c**, Dot plots indicating IC_{80} (μg ml^{-1}) of 3BNC117 (**b**) and 10-1074 (**c**) against latent and rebound viruses determined by TZM-bl neutralization assay. Q^2VOA-derived latent viruses from week −2 and week 12 are shown as black and grey circles, respectively. For outgrowth culture-derived rebound viruses, the highest IC_{80} is shown as red circle. For 9246, 9252, 9245 and 9251 viruses could not be obtained from rebound outgrowth cultures and pseudoviruses were made from *env* sequences from Q^2VOA and plasma SGA.

suppression for at least 15 weeks, although two rebounded at weeks 5 and 7 (Fig. 1b, c).

Quantitative and qualitative viral outgrowth assays (Q^2VOA) were used to retrospectively analyse the replication-competent latent viral reservoir in all individuals. Phylogenetic analysis showed that the trial participants were infected with epidemiologically distinct clade B viruses (Extended Data Fig. 3). Q^2VOA analysis revealed that the pre-infusion latent reservoir in the two individuals who rebounded early, 9245 and 9251, harboured 10-1074- or 3BNC117-resistant viruses, respectively (Fig. 2 and Supplementary Table 4). Therefore, these two individuals were effectively subjected to antibody monotherapy, because there was pre-existing resistance in the reservoir of these individuals to one of the two bNAbs. Consistent with this idea, the delay in rebound in these two participants was within the anticipated range of antibody monotherapy[9,11] (Fig. 1c). In addition, all four of the individuals excluded from the analysis due to incomplete viral suppression showed pre-existing resistance or viruses that were not fully neutralized by one or both of the antibodies and these individuals rebounded before week 12 (Extended Data Figs. 4, 5 and Supplementary Table 4).

To examine the viruses that arose in the early rebounding individuals, we performed single genome analysis (SGA) of plasma viruses obtained at the time of rebound. In addition to the pre-existing sequences associated with resistance in the 10-1074 target site (N332T and S334N, Fig. 2a), rebound viruses in 9245 also carried an extended V5 loop and potential *N*-linked glycosylation sites that could interfere with 3BNC117 binding (Extended Data Fig. 6). Conversely, genetic features associated with resistance to 3BNC117 were found in the pre-infusion reservoir of 9251 and were accompanied by mutations in the 10-1074 target site in the rebounding viruses (S334N, Fig. 2a and Extended Data Fig. 6). For both individuals, resistance of rebound viruses to both antibodies was confirmed by the TZM-bl neutralization assay (Fig. 2b, c and Supplementary Table 4). Thus, bulk outgrowth cultures used for screening failed to detect pre-existing resistance in the reservoir of 2 of the 11 studied individuals. This result is not surprising given that bulk cultures are dominated by a limited number of rapidly growing viral species that may not be representative of the diversity of the latent reservoir.

The median time to rebound in the seven individuals that had no detectable resistant viruses in the pre-infusion latent reservoir, and rebounded during the study period, was also 21 weeks and different from the 6–10 weeks found for monotherapy with 3BNC117[9] (Fig. 1c). In these participants, viral suppression was maintained for 15–26 weeks after ART discontinuation (Supplementary Table 2). The two remaining participants (9254 and 9255) completed study follow-up at 30 weeks without experiencing rebound (Supplementary Table 2). Notably, viral rebound never occurred when the concentration of both administered antibodies was above 10 μg ml^{-1}. The average serum concentration of 3BNC117 (determined by TZM-bl assay) at the time of rebound in sensitive individuals that rebounded during study follow-up was 1.9 μg ml^{-1} (Fig. 1b and Supplementary Table 2). By contrast, the average serum concentration of 10-1074 at rebound was 14.8 μg ml^{-1} (Fig. 1b and Supplementary Table 2). The difference in the antibody concentrations at the time of rebound is consistent with the longer half-life of 10-1074, which resulted in a period of 10-1074 monotherapy (Fig. 1b, Extended Data Fig. 2c–f and Supplementary Table 2). Finally, these nine individuals showed little or no pre-existing neutralizing antibodies against a diagnostic panel of viruses before bNAb infusion (Supplementary Table 5).

© 2018 Springer Nature Limited. All rights reserved.

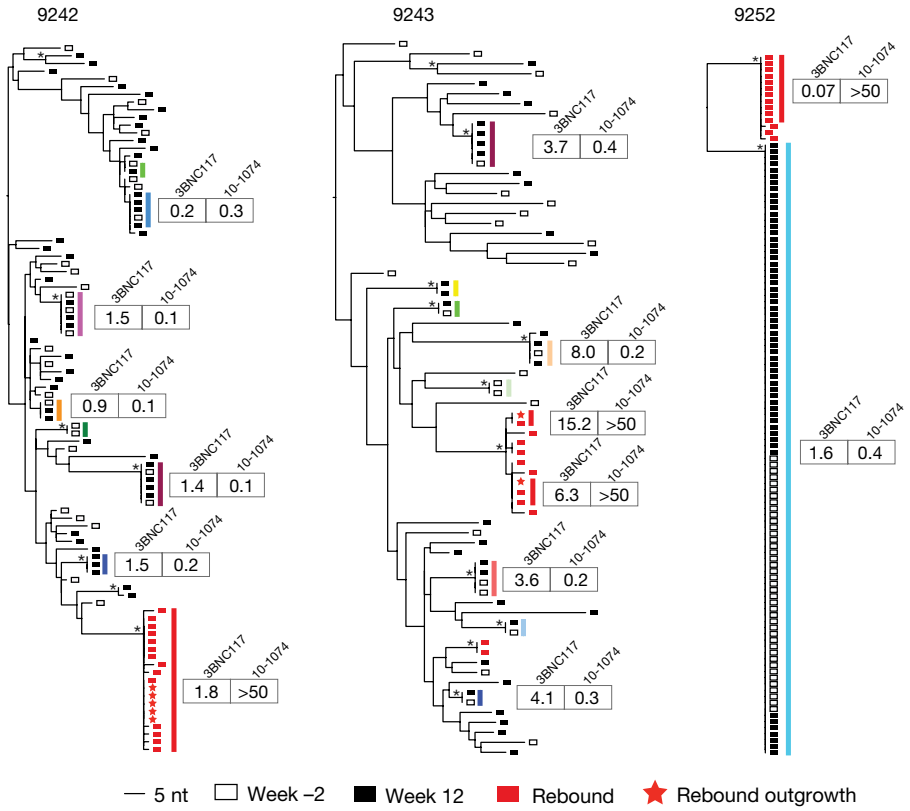


Fig. 3 | Comparison of the circulating latent reservoir and rebound viruses. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q^2VOA, rebound plasma SGA and rebound PBMC outgrowth cultures from three out of seven participants (9242, 9243 and 9252) that rebounded before week 30 (9241, 9244, 9247 and 9246 are depicted in Extended Data Fig. 7). Open and closed black rectangles indicate Q^2VOA-derived viruses from week −2 and week 12, respectively. Viruses obtained at the time of rebound are indicated by red rectangles (plasma SGA) and red stars (rebound PBMC outgrowth cultures). Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥70%). Clones are denoted by coloured lines mirroring the colours of slices in Extended Data Fig. 10a. Boxes indicate IC_{80} values (μg ml^{-1}) of 3BNC117 and 10-1074 against representative viruses throughout the phylogenetic tree and clones, when possible (Supplementary Table 4). nt, nucleotide.

Rebound and latent viruses

To examine the relationship between rebound viruses and the circulating latent reservoir, we compared *env* sequences obtained from plasma rebound viruses by SGA with sequences obtained by Q^2VOA from both pre-infusion and week 12 samples. In addition, we measured the sensitivity of rebound outgrowth viruses and/or pseudoviruses to 3BNC117 and 10-1074 using the TZM-bl neutralization assay (Fig. 2b, c, 3, Extended Data Fig. 7 and Supplementary Table 4). A total of 154 viral *env* sequences obtained by plasma SGA were analysed and compared to 408 sequences obtained from the latent reservoir by Q^2VOA. Although rebound and reservoir viruses clustered together for each individual (Extended Data Fig. 3), we found no identical sequences between the two compartments in any of the individuals studied (Figs. 3, 4a and Extended Data Fig. 7). The difference could be accounted for by distinct requirements for HIV-1 reactivation in vitro and in vivo, compartmentalization of reservoir viruses, HIV-1 mutation during the course of the trial and/or by viral recombination in some individuals[20,21] (Extended Data Fig. 8). Whether or not bNAb therapy influences selection for recombination events remains to be determined.

Similar to 3BNC117 monotherapy, the vast majority of rebounding viruses clustered within low-diversity lineages consistent with expansion of 1–2 recrudescent viruses[9] (Fig. 3, Extended Data Figs. 7, 9). By contrast, rebound viruses are consistently polyclonal during ATI in the absence of antibody therapy[22,23]. Thus, the antibodies restrict the outgrowth of latent viruses in vivo.

The emerging viruses in 6 of the 7 individuals who rebounded when the mean 3BNC117 and 10-1074 serum concentrations were 1.9 and 14.8 μg ml^{-1}, respectively, carried resistance-associated mutations in the 10-1074 target site (Figs. 1b, 2a). Consistent with the sequence data, these rebound viruses were generally resistant to 10-1074, as shown by the TZM-bl neutralization assay, but remained sensitive to 3BNC117 (Fig. 2b, c and Supplementary Table 4). The level of sensitivity to 3BNC117 in these emerging viruses was similar to that found in the reservoir viruses in each of the individuals (Fig. 2b and Supplementary Table 4). One individual, 9244, showed rebound viruses that remained sensitive to both antibodies in TZM-bl neutralization assays. Rebound occurred when 3BNC117 and 10-1074 concentrations in serum of this individual were undetectable and 11.6 μg ml^{-1}, respectively (Fig. 1b and Supplementary Table 2). The sensitivity of the plasma rebound viruses was similar to that of latent pre-infusion and week 12 viruses obtained in viral outgrowth cultures (Fig. 2b, c and Supplementary Table 4). Therefore, this individual did not develop resistance to either of the antibodies despite prolonged exposure to both. In conclusion, none of the nine individuals with pre-infusion reservoirs containing viruses that were sensitive to both antibodies developed double resistance during the observation period.

The latent reservoir

To determine whether there were changes in the circulating reservoir during the observation period, we compared the results of Q^2VOA assays performed at entry and 12 weeks after the start of ATI for 8 of the 9 individuals that remained suppressed for at least 12 weeks (Fig. 4 and Extended Data Fig. 10). Similar to previous reports, 63% of all viruses obtained by Q^2VOA belonged to expanded clones[20,24–26] (Extended Data Fig. 10a, b). Comparison of the *env* sequences of the viruses that emerged in outgrowth cultures revealed that 60% of the sequences could be found at both time points. However, there were numerous examples of clones that appeared or disappeared between

© 2018 Springer Nature Limited. All rights reserved.

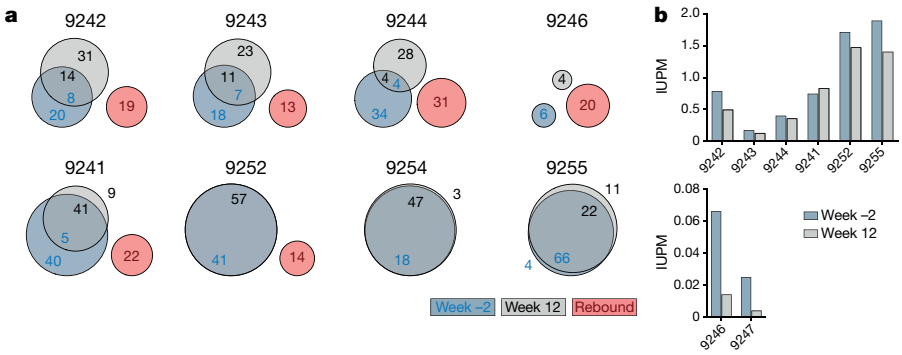


Fig. 4 | Distribution of the circulating latent reservoir and rebound viruses. a, Venn diagrams showing sequence identity between *env* sequences obtained from Q^2VOA at week −2 (blue) and week 12 (grey), and plasma SGA or rebound PBMC outgrowth culture at the time of viral rebound (red). Area of overlap is proportional to the number of identical sequences. The number of obtained sequences is indicated. **b**, IUPM $CD4^+$ T cells at weeks −2 and 12 as determined by Q^2VOA. Participants with IUPMs that were higher and lower than 0.1 are shown at the top and bottom, respectively. Participant 9254 is not shown owing to lack of sample availability. The two time points were not statistically different ($P = 0.078$ (paired Student's *t*-test)).

the time points and some of the changes were significant (Extended Data Fig. 10a). To determine the number of infectious units per million (IUPM, http://silicianolab.johnshopkins.edu/), 6.0×10^7–6.2×10^8 $CD4^+$ T cells were assayed by Q^2VOA for each time point for each individual (Fig. 4b). The difference between the two time points was never greater than 6.5-fold for any individual, and the IUPM values at the two time points were not statistically different ($P = 0.078$). Moreover, time to rebound was not directly correlated with IUPM (Extended Data Fig. 10c). Additional time points would be required to calculate the half-life of the reservoir in individuals who received immunotherapy[27].

Discussion

First-generation anti-HIV-1 bNAbs were generally ineffective in suppressing viraemia in animal models and humans leading to the conclusion that this approach should not be pursued[17,18,28]. The advent of new methods for anti-HIV-1 antibody cloning[29] and subsequent discovery of a new, more potent generation of bNAbs revitalized this area of research[30,31].

bNAb monotherapy with 3BNC117 or VRC01 is not enough to maintain control during ATI in HIV-1-infected humans[9,11]. Similar results were obtained in participant 9251 who effectively received 10-1074 monotherapy due to pre-existing resistance to 3BNC117. By contrast the combination of 3BNC117 and 10-1074 is sufficient to maintain viral suppression in sensitive individuals when the concentration of both antibodies remains above of 10 μg ml^{-1} in serum. Rebound occurred when 3BNC117 levels dropped below 10 μg ml^{-1} effectively leading to 10-1074 monotherapy, from which viruses in nearly all individuals rapidly escaped by mutations in the 10-1074 contact site. The observation that nine individuals infected with distinct viruses were unable to develop viruses who were resistant to both antibodies over a median period of 21 weeks suggests that viral replication was severely limited by this combination of antibodies.

In human studies, monotherapy with 3BNC117 is associated with enhanced humoral immunity and accelerated clearance of HIV-1-infected cells[5,32]. In addition, when administered early to macaques infected with the chimeric simian/human immunodeficiency virus $SHIV_{AD8}$, combined 3BNC117 and 10-1074 immunotherapy induced host $CD8^+$ T cell responses that contributed to the control of viraemia in nearly 50% of the animals[3]. However, virus-specific $CD8^+$ T cells that were responsible for control of viraemia in these macaques were not detected in the circulation, and their contribution to viral suppression was only documented after $CD8^+$ T cell depletion[3]. In most macaques that maintained viral control, complete viral suppression was only established after rebound viraemia that followed antibody clearance[3].

Two individuals in this study remained suppressed for over 30 weeks after ATI, 9254 and 9255. Neither participant had detectable levels of ART in the blood or carried the B*27 and B*57 HLA alleles that are most frequently associated with elite control[33]. The first, 9254, reports starting ART within 4–5 months after probable exposure to the virus with an initial viral load of 860,000 copies per ml. Despite relatively early therapy, and excellent virological control for 21 years on therapy, this individual had an IUPM of 0.68 by Q^2VOA at the 12-week time point (Extended Data Fig. 10b). The second individual, 9255, showed several viral blips that were spontaneously controlled beginning 15 weeks after ATI when antibody levels were waning. This individual was infected for at least 7 months before starting ART with an initial viral load of 85,800 copies per ml and had an IUPM of 1.4 at the 12-week time point. A small fraction of individuals on ART[10] show spontaneous prolonged virologic control after ART is discontinued, and this number appears to increase when ART treatment is initiated during the acute phase of infection[34–38]. Whether antibody-enhanced $CD8^+$ T cell responses contribute to the prolonged control in the two out of nine individuals who received combination immunotherapy and whether this effect can be enhanced by latency reactivating agents or immune checkpoint inhibitors remains to be determined.

A substantial fraction of the circulating latent reservoir is composed of expanded clones of infected T cells[24,26,39–42]. These T cell clones appear to be dynamic in that the specific contribution of individual clones of circulating latently infected $CD4^+$ T cells to the reservoir of individuals receiving ART fluctuates over time[24,25]. Individuals that maintain viral suppression by antibody therapy appear to show similar fluctuations in reservoir clones that do not appear to be associated with antibody sensitivity. Whether the apparent differences observed in the reservoir during immunotherapy lead to changes in the reservoir half-life cannot be determined from the available data and will require reservoir assessments in additional individuals at multiple time points over an extended observation period.

Individuals harbouring viruses sensitive to 3BNC117 and 10-1074 maintained viral suppression during ATI for a median of almost four months after the final antibody administration. However, HIV-1 is a highly diverse virus with varying levels of sensitivity to specific bNAbs. As a result, maintenance therapy with just the combination of 3BNC117 and 10-1074 would only be possible for the approximately 50% of clade B-infected individuals that are sensitive to both antibodies. This problem may be overcome by addition of or substitution with other antibodies[14,15,43], or long-acting small-molecule antiretroviral drugs.

In macaques, the therapeutic efficacy of anti-HIV-1 antibodies is directly related to their half-life[4,12,13], which can be extended by mutations that enhance Fc domain interactions with the neonatal Fc receptor[4,13,44]. These mutations also increase the half-life of antibodies in humans by 2–4-fold[45]. Our data suggest that a single administration of combinations of bNAbs with extended half-lives could maintain suppression for 6–12 months in individuals harbouring sensitive viruses.

© 2018 Springer Nature Limited. All rights reserved.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0531-2.

Received: 6 June 2018; Accepted: 30 July 2018;
Published online 26 September 2018.

1. Doitsh, G. & Greene, W. C. Dissecting how CD4 T cells are lost during HIV infection. *Cell Host Microbe* **19**, 280–291 (2016).
2. Churchill, M. J., Deeks, S. G., Margolis, D. M., Siliciano, R. F. & Swanstrom, R. HIV reservoirs: what, where and how to target them. *Nat. Rev. Microbiol*. **14**, 55–60 (2016).
3. Nishimura, Y. et al. Early antibody therapy can induce long-lasting immunity to SHIV. *Nature* **543**, 559–563 (2017).
4. Gautam, R. et al. A single injection of crystallizable fragment domain-modified antibodies elicits durable protection from SHIV infection. *Nat. Med*. **24**, 610–616 (2018).
5. Schoofs, T. et al. HIV-1 therapy with monoclonal antibody 3BNC117 elicits host immune responses against HIV-1. *Science* **352**, 997–1001 (2016).
6. Caskey, M. et al. Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522**, 487–491 (2015).
7. Lynch, R. M. et al. Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. *Sci. Transl. Med*. **7**, 319ra206 (2015).
8. Caskey, M. et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nat. Med*. **23**, 185–191 (2017).
9. Scheid, J. F. et al. HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature* **535**, 556–560 (2016).
10. Li, J. Z. et al. The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. *AIDS* **30**, 343–353 (2016).
11. Bar, K. J. et al. Effect of HIV antibody VRC01 on viral rebound after treatment interruption. *N. Engl. J. Med*. **375**, 2037–2050 (2016).
12. Shingai, M. et al. Passive transfer of modest titers of potent and broadly neutralizing anti-HIV monoclonal antibodies block SHIV infection in macaques. *J. Exp. Med*. **211**, 2061–2074 (2014).
13. Gautam, R. et al. A single injection of anti-HIV-1 antibodies protects against repeated SHIV challenges. *Nature* **533**, 105–109 (2016).
14. Klein, F. et al. HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* **492**, 118–122 (2012).
15. Horwitz, J. A. et al. HIV-1 suppression and durable control by combining single broadly neutralizing antibodies and antiretroviral drugs in humanized mice. *Proc. Natl Acad. Sci. USA* **110**, 16538–16543 (2013).
16. Shingai, M. et al. Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503**, 277–280 (2013).
17. Trkola, A. et al. Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. *Nat. Med*. **11**, 615–622 (2005).
18. Mehandru, S. et al. Adjunctive passive immunotherapy in human immunodeficiency virus type 1-infected individuals treated with antiviral therapy during acute and early infection. *J. Virol*. **81**, 11016–11031 (2007).
19. Sarzotti-Kelsoe, M. et al. Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *J. Immunol. Methods* **409**, 131–146 (2014).
20. Cohen, Y. Z. et al. Relationship between latent and rebound viruses in a clinical trial of anti–HIV-1 antibody 3BNC117. *J. Exp. Med*. **215**, https://doi.org/10.1084/jem.20180936 (2018).
21. Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. Recombination in HIV-1. *Nature* **374**, 124–126 (1995).
22. Rothenberger, M. K. et al. Large number of rebounding/founder HIV variants emerge from multifocal infection in lymphatic tissues after treatment interruption. *Proc. Natl Acad. Sci. USA* **112**, E1126–E1134 (2015).
23. Kearney, M. F. et al. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog*. **10**, e1004010 (2014).
24. Lorenzi, J. C. et al. Paired quantitative and qualitative assessment of the replication-competent HIV-1 reservoir and comparison with integrated proviral DNA. *Proc. Natl Acad. Sci. USA* **113**, E7908–E7916 (2016).
25. Wang, Z. et al. Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc. Natl Acad. Sci. USA* **115**, E2575–E2584 (2018).
26. Hosmane, N. N. et al. Proliferation of latently infected $CD4^+$ T cells carrying replication-competent HIV-1: potential role in latent reservoir dynamics. *J. Exp. Med*. **214**, 959–972 (2017).
27. Crooks, A. M. et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J. Infect. Dis*. **212**, 1361–1365 (2015).
28. Poignard, P. et al. Neutralizing antibodies have limited effects on the control of established HIV-1 infection in vivo. *Immunity* **10**, 431–438 (1999).
29. Scheid, J. F. et al. A method for identification of HIV gp140 binding memory B cells in human blood. *J. Immunol. Methods* **343**, 65–67 (2009).
30. Escolano, A., Dosenovic, P. & Nussenzweig, M. C. Progress toward active or passive HIV-1 vaccination. *J. Exp. Med*. **214**, 3–16 (2017).
31. Kwong, P. D. & Mascola, J. R. HIV-1 vaccines based on antibody identification, B Cell ontogeny, and epitope structure. *Immunity* **48**, 855–871 (2018).
32. Lu, C. L. et al. Enhanced clearance of HIV-1-infected cells by broadly neutralizing antibodies against HIV-1 in vivo. *Science* **352**, 1001–1004 (2016).
33. Walker, B. D. & Yu, X. G. Unravelling the mechanisms of durable control of HIV-1. *Nat. Rev. Immunol*. **13**, 487–498 (2013).
34. Colby, D. J. et al. Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat. Med*. **24**, 923–926 (2018).
35. Sáez-Cirión, A. et al. Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathog*. **9**, e1003211 (2013).
36. Sneller, M. C. et al. A randomized controlled safety/efficacy trial of therapeutic vaccination in HIV-infected individuals who initiated antiretroviral therapy early in infection. *Sci. Transl. Med*. **9**, eaan8848 (2017).
37. Fidler, S. et al. Virological blips and predictors of post treatment viral control after stopping ART started in primary HIV infection. *J. Acquir. Immune Defic. Syndr*. **74**, 126–133 (2017).
38. Martin, G. E. et al. Post-treatment control or treated controllers? Viral remission in treated and untreated primary HIV infection. *AIDS* **31**, 477–484 (2017).
39. Cohn, L. B. et al. Clonal $CD4^+$ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med*. **24**, 604–609 (2018).
40. Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).
41. Wagner, T. A. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
42. Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
43. Halper-Stromberg, A. et al. Broadly neutralizing antibodies and viral inducers decrease rebound from HIV-1 latent reservoirs in humanized mice. *Cell* **158**, 989–999 (2014).
44. Ko, S. Y. et al. Enhanced neonatal Fc receptor function improves protection against primate SHIV infection. *Nature* **514**, 642–645 (2014).
45. Gaudinski, M. R. et al. Safety and pharmacokinetics of the Fc-modified HIV-1 human monoclonal antibody VRC01LS: a phase 1 open-label clinical trial in healthy adults. *PLoS Med*. **15**, e1002493 (2018).

Acknowledgements We thank all study participants who devoted time to our research; members of the Klein and Nussenzweig laboratories for helpful discussions, especially Y. Bar-On, L. Cohn and M. Jankovic; R. Levin for study coordination and the Rockefeller University Hospital Clinical Research Support Office and nursing staff as well as K. Fiddike, C. Golder, S. Margane, M. Platten, E. Voigt and D. Weiland for help with recruitment and study implementation; K. Jain for help with sample processing; S. Kiss for ophthalmologic assessments; T. Keler and the Celldex Therapeutics team for 3BNC117 and 10-1074 manufacturing and regulatory support; C. Conrad for regulatory support; U. Kerkweg, R. Macarthur and A. Johnson for pharmaceutical services; H. Janicki, M. Ercanoglu, P. Schommers and R. Kaiser for help with virus cultures; C. Scheid and U. Holtick for leukaphereses; S. McMillan, S. Mosher, S. Sawant, D. Beaumont, M. Sarzotti-Kelsoe, K. Greene, H. Gao and D. Montefiori for help with PK assay development, validation, reporting, and/or project management; P. Fast and H. Park for clinical monitoring; and S. Schlesinger for input on study design. This work was supported by the Bill and Melinda Gates Foundation Collaboration for AIDS Vaccine Discovery (CAVD) grants OPP1092074, OPP1124068 (M.C.N.), CAVIMC OPP1146996 (G.D.T., M.S.S.); the Heisenberg-Program of the DFG (KL 2389/2-1), the European Research Council (ERC-StG639961), and the German Center for Infection Research (DZIF) (F.K.); the NIH grants 1UM1 AI100663 and R01AI-129795 (M.C.N.); the Einstein-Rockefeller-CUNY Center for AIDS Research (1P30AI124414-01A1); BEAT-HIV Delaney grant UM1 AI126620 (M.C.); and the Robertson fund of the Rockefeller University. M.C.N. is a Howard Hughes Medical Institute Investigator.

Reviewer information *Nature* thanks G. Silvestri and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.C. is the principal investigator for the work in the United States and F.K. is the principal investigator for Germany; M.C., F.K. and M.C.N. designed the trial; P.M., H.G., F.K., M.C. and M.C.N. analysed the data and wrote the manuscript; P.M., L.N. and T.Ka. performed Q^2VOA, rebound cultures and SGA; H.G., M.W.-P., G.K., E.T., J.H., M.C. and F.K. implemented the study; A.L.B., K.M., Y.Z.C., C.L., I.Su., C.W., T.Kü. and C.S. contributed to recruitment and clinical assessments; P.M., H.G. and L.N. performed bulk viral cultures; J.A.P. and T.Y.O. performed bioinformatics processing; K.E.S. and G.D.T. conducted anti-idiotype ELISA; M.S.S. conducted TZM-bl assays; C.U.-O., R.P., C.R., M.S. and I.Sh. coordinated and performed sample processing; L.H., A.P.W., P.J.B. and N.P. contributed to data analysis; J.C.C.L., C-L.L., R.M.G. and G.F. contributed to study design and implementation.

Competing interests : There are patents on 3BNC117 (PTC/US2012/038400) and 10-1074 (PTC/US2013/065696) that list M.C.N. as an inventor.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0531-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0531-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.K., M.C. or M.C.N.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Study design. An open-label phase 1b study was conducted in HIV-1-infected participants who were virologically suppressed on ART (http://www.clinicaltrials.gov; NCT02825797; EudraCT: 2016-002803-25). Study participants were enrolled sequentially according to eligibility criteria. Participants received 3BNC117 and 10-1074 intravenously at a dose of 30 mg kg^{-1} body weight of each antibody, at weeks 0, 3 and 6, unless viral rebound occurred. ART was discontinued 2 days after the first infusion of antibodies (day 2). Plasma HIV-1 viral RNA levels were monitored weekly and ART was resumed if the viral load increased to $\geq$200 copies per ml or $CD4^+$ T cell counts decreased to <350 cells per μl in two consecutive measurements. Time to viral rebound was determined by the first of two consecutive viral loads of >200 copies per ml. Study participants were followed for 30 weeks after the first infusion. Safety data are reported until the end of study follow-up. All participants provided written informed consent before participation in the study and the study was conducted in accordance with Good Clinical Practice. The protocol was approved by the Federal Drug Administration in the USA, the Paul-Ehrlich-Institute in Germany, and the Institutional Review Boards (IRBs) at the Rockefeller University and the University of Cologne.

Study participants. Study participants were recruited at the Rockefeller University Hospital, New York, USA, and the University Hospital Cologne, Cologne, Germany. Eligible participants were adults aged 18–65 years, HIV-1-infected, on ART for a minimum of 24 months, with plasma HIV-1 RNA levels of <50 copies per ml for at least 18 months (one viral blip of >50 but <500 copies per ml during this 18-month period was allowed), plasma HIV-1 RNA levels <20 copies per ml at the screening visit, and a current $CD4^+$ T cell count >500 cells per μl. In addition, participants were prescreened for sensitivity of latent proviruses against 3BNC117 and 10-1074 by bulk PBMC viral outgrowth culture as described in 'Prescreening bulk PBMC cultures'. Sensitivity was defined as an $IC_{50} < 2$ μg ml^{-1} for both 3BNC117 and 10-1074 against outgrowth virus. Participants on an ART regimen that included a non-nucleoside reverse transcriptase inhibitor (NNRTI) were switched to an integrase inhibitor-based regimen (dolutegravir plus tenofovir disoproxil fumarate and emtricitabine) four weeks before treatment interruption due to the prolonged half-life of NNRTIs. Exclusion criteria included reported $CD4^+$ T cell nadir of <200 cells μl^{-1}, concomitant hepatitis B or C infection, previous receipt of monoclonal antibodies of any kind, clinically relevant physical findings, medical conditions or laboratory abnormalities, and pregnancy or lactation.

Study procedures. 3BNC117 and 10-1074 were administered intravenously at a dose of 30 mg kg^{-1}. The appropriate stock volume of 3BNC117 and 10-1074 was calculated according to body weight and diluted in sterile normal saline to a total volume of 250 ml per antibody. Monoclonal antibody infusions were administered sequentially and intravenously over 60 min. Study participants were observed at the Rockefeller University Hospital or the University Hospital Cologne for 1 h after the last antibody infusion. Participants returned for weekly follow-up visits during the ATI period for safety assessments, which included physical examination and measurements of clinical laboratory parameters such as haematology, chemistries, urinalysis and pregnancy tests (for women). Plasma HIV-1 RNA levels were monitored weekly during the ATI period and $CD4^+$ T cell counts were measured every 1–2 weeks. After ART was re-initiated, participants returned for follow-up every two weeks until viral re-suppression was achieved, and every eight weeks thereafter. Study investigators evaluated and graded adverse events according to the Division of AIDS (DAIDS) Table for Grading the Severity of Adult and Pediatric Adverse Events (version 2.0, November 2014) and determined causality. Leukapheresis was performed at the Rockefeller University Hospital or at the University Hospital Cologne at week –2 and week 12. Blood samples were collected before and at multiple times after 3BNC117 and 10-1074 infusions. Samples were processed within 4 h of collection, and serum and plasma samples were stored at −80 °C. PBMCs were isolated by density gradient centrifugation. The absolute number of PBMCs was determined using an automated cell counter (Vi-Cell XR; Beckman Coulter) or manually, and cells were cryopreserved in fetal bovine serum plus 10% DMSO.

Plasma HIV-1 RNA Levels. HIV-1 RNA levels in plasma were measured at the time of screening, at week –2, day 0 (before infusion), weekly during ATI, and every two weeks to every eight weeks after viral rebound had occurred. HIV-1 RNA levels were determined using the Roche COBAS AmpliPrep/COBAS TaqMan HIV-1 Assay (version 2.0) or the Roche COBAS HIV-1 quantitative nucleic acid test (COBAS 6800), which quantify HIV-1 RNA over a range of 2×10^1 to 1×10^7 copies per ml. These assays were performed at LabCorp or at the University Hospital Cologne.

$CD4^+$ T cells. $CD4^+$ T cell counts were determined by clinical flow cytometry assay, performed at LabCorp or at the University Hospital Cologne, at screening, week 0 (before infusion), weeks 2, 3, 5, 6, 8, 10, and weekly thereafter, while participants remained off ART.

Determination of baseline neutralizing antibody activity. Purified IgG (Protein G Sepharose 4 Fast Flow, GE Life Sciences) obtained before antibody infusions was tested against a panel of 12 HIV-1 pseudoviruses as described previously[5].

Measurement of 3BNC117 and 10-1074 serum levels. Blood samples were collected before, at the end of each 3BNC117 infusion and at the end of each 10-1074 infusion at weeks 0, 3 and 6, and weekly during the ATI period, up to week 30. Serum levels of 3BNC117 and 10-1074 were determined by a TZM-bl assay and by ELISA from samples obtained before and after each antibody infusion, and approximately every three weeks during follow-up as well as at the time of viral rebound.

Serum concentrations of 3BNC117 and 10-1074 were measured by a validated sandwich ELISA. High-bind polystyrene plates were coated with 4 μg ml^{-1} of an anti-idiotypic antibody that specifically recognizes 3BNC117 (anti-ID 1F1-2E3 monoclonal antibody) or 2 μg ml^{-1} of an anti-idiotypic antibody that specifically recognizes 10-1074 (anti-ID 3A1-4E11 monoclonal antibody), and incubated overnight at 2–8 °C. After washing, plates were blocked with 5% Milk Blotto (w/v), 5% NGS (v/v) and 0.05% Tween 20 (v/v) in PBS. Serum samples, quality controls and standards were added (1:50 minimum dilution in 5% Milk Blotto (w/v), 5% NGS (v/v) and 0.05% Tween 20 (v/v) in PBS) and incubated at room temperature. 3BNC117 or 10-1074 were detected using a horseradish peroxidase (HRP)-conjugated mouse anti-human IgG kappa-chain-specific antibody (Abcam) for 3BNC117 or an HRP-conjugated goat anti-human IgG Fc-specific antibody for 10-1074 (Jackson ImmunoResearch) and the HRP substrate tetra-methylbenzidine. 3BNC117 and 10-1074 concentrations were then calculated from the standard curves of 3BNC117 or 10-1074 that were run on the same plate using a 5-PL curve-fitting algorithm (Softmax Pro, v.5.4.5). Standard curves and positive controls were created from the drug product lots of 3BNC117 and 10-1074 used in the clinical study. The capture anti-idiotypic monoclonal antibodies were produced using a stable hybridoma cell line (Duke Protein Production Facility[6]). The lower limit of quantification for the 3BNC117 ELISA is 0.78 μg ml^{-1} and for the 10-1074 ELISA is 0.41 μg ml^{-1}. The lower limit of detection was determined to be 0.51 μg ml^{-1} and 0.14 μg ml^{-1} in HIV-1 seropositive serum for the 3BNC117 and 10-1074 ELISA, respectively. For values that were detectable (that is, positive for the monoclonal antibodies) but were below the lower limit of quantification, values are reported as <0.78 μg ml^{-1} and <0.41 μg ml^{-1} for 3BNC117 and 10-1074 ELISA, respectively. If day 0 baseline samples had measurable levels of antibody by the respective assays, the background measured antibody level was subtracted from subsequent results. In addition, samples with antibody levels measured to be within threefold from background were excluded from the analysis of pharmacokinetic parameters.

Serum concentrations of active 3BNC117 and 10-1074 were also measured using a validated luciferase-based neutralization assay in TZM-bl cells as previously described[19]. In brief, serum samples were tested using a primary 1:20 dilution with a fivefold titration series against HIV-1 Env pseudoviruses Q769.d22 and X2088_c9, which are highly sensitive to neutralization by 3BNC117 and 10-1074, respectively, while fully resistant against the other administered antibody. In the case of the post-infusion time points of 10-1074, instances for which the serum 50% inhibitory dilution (ID_{50}) titres against X2088_c9 were >100,000, serum samples were also tested against a less sensitive strain, Du422 (Supplementary Table 2). To generate standard curves, clinical drug products of 3BNC117 and 10-1074 were included in every assay set-up using a primary concentration of 10 μg ml^{-1} with a fivefold titration series. Serum concentrations of 3BNC117 and 10-1074 for each sample were calculated as follows: serum ID_{50} titre (dilution) × 3BNC117 IC_{50} or 10-1074 IC_{50} titre (μg ml^{-1}) = serum concentration of 3BNC117 or 10-1074 (μg ml^{-1}). Env pseudoviruses were produced using an ART-resistant backbone vector that reduces background inhibitory activity of antiretroviral drugs if present in the serum sample (SG3ΔEnv/K101P.Q148H.Y181C; M.S.S., unpublished data). Virus that was pseudotyped with the envelope protein of murine leukaemia virus (MuLV) was used as a negative control. Antibody concentrations were calculated using the serum ID_{80} titre and monoclonal antibody IC_{80} if non-specific activity against MuLV was detected ($ID_{50} > 20$; 9246, week 30; 9248, baseline, day 0, week 18). All assays were performed in a laboratory that meets Good Clinical Laboratory Practice standards.

Prescreening bulk PBMC cultures. To test HIV-1 viral strains for sensitivity to 3BNC117 and 10-1074, we performed bulk viral outgrowth cultures by coculturing isolated $CD4^+$ T cells with the MOLT-4/CCR-5 cell line (NIH AIDS Reagent Program, Ca. No. 4984) or $CD8^+$ T cell-depleted healthy donor lymphoblasts. PBMCs for prescreening were obtained up to 72 weeks (range 54–505 days) before enrollment under separate protocols approved by the IRBs of the Rockefeller University and the University of Cologne. Sensitivity was determined by TZM-bl neutralization assay as described below. Culture supernatants with $IC_{50} < 2$ μg ml^{-1} were deemed sensitive.

Quantitative and qualitative viral outgrowth assay. The Q^2VOA was performed using isolated PBMCs from leukapheresis at week –2 and week 12 as previously described[24]. In brief, isolated $CD4^+$ T cells were activated with 1 μg ml^{-1} phytohaemagglutinin (PHA; Life Technologies) and 100 U ml^{-1} IL-2 (Peprotech) and cocultured with 1×10^6 irradiated PBMCs from a healthy donor in 24-well plates.

© 2018 Springer Nature Limited. All rights reserved.

A total of 6×10^7–6.2×10^8 cells were assayed for each individual at each of the two time points. After 24 h, PHA was removed and 0.1×10^6 MOLT-4/CCR5 cells were added to each well. Cultures were maintained for two weeks, splitting the MOLT- 4/CCR5 cells in half seven days after the initiation of the culture and every other day after that. Positive wells were detected by measuring p24 by ELISA. The frequency of latently infected cells was calculated through the IUPM algorithm developed by the Siliciano laboratory (http://silicianolab.johnshopkins.edu).

Rebound outgrowth cultures. $CD4^+$ T cells isolated from PBMCs from the rebound time points were cultured at limiting dilution exactly as described for Q^2VOA. $CD4^+$ T cells were activated with T cell activation beads (Miltenyi) at a concentration of 0.5×10^6 beads per 10^6 $CD4^+$ T cells and 20 U ml^{-1} of IL-2. Rebound outgrowth cultures were performed using PBMCs from the highest viral load sample (usually the repeat measurement $\geq$200 copies per ml). Viruses for which the sequences matched the SGA *env* sequences, and therefore were identical to those present in plasma, as opposed to potentially reactivated PBMC-derived latent reservoir viruses, were selected to test for neutralization.

Viral sensitivity testing. Supernatants from p24-positive bulk PBMC cultures, rebound PBMC outgrowth cultures and Q^2VOA wells were tested for sensitivity to 3BNC117 and 10-1074 by TZM-bl neutralization assay as previously described[19].

Sequencing. HIV-1 RNA extraction and single-genome amplification was performed as previously described[46]. In brief, HIV-1 RNA was extracted from plasma samples or Q^2VOA-derived virus supernatants using the MinElute Virus Spin kit (Qiagen) followed by first-strand cDNA synthesis using SuperScript III reverse transcriptase (Invitrogen). cDNA synthesis for plasma-derived HIV-1 RNA was performed using the antisense primer envB3out 5′-TTGCTACTTGTGATTGCTCCATGT-3′. gp160 was amplified using envB5out 5′-TAGAGCCCTGGAAGCATCCAGGAAG-3′ and envB3out 5′-TTGCTACTTGTGATTGCTCCATGT-3′ in the first round and in the second round with nested primers envB5in 5′-CACCTTAGGCATCTCCTATGGCAGGAAGAAG-3′ and envB3in 5′-GTCTCGAGATACTGCTCCCACCC-3′. PCRs were performed using High Fidelity Platinum Taq (Invitrogen) and run at 94 °C for 2 min; 35 cycles of 94 °C for 15 s, 55 °C for 30 s and 68 °C for 4 min; and 68 °C for 15 min. Second-round PCR was performed with 1 μl of the PCR product from the first round as template and High Fidelity Platinum Taq at 94 °C for 2 min; 45 cycles of 94 °C for 15 s, 55 °C for 30 s and 68 °C for 4 min; and 68 °C for 15 min. cDNA synthesis for Q^2VOA-derived HIV-1 RNA was performed using the antisense primer R3B6R 5′-TGAAGCACTCAAGGCAAGCTTTATTGAGGC-3′. The *env* 3′ half-genome was amplified in a single PCR using B3F3 primer 5′-TGGAAAGGTGAAGGGGCAGTAGTAATAC-3′ and R3B6R primer 5′-TGAAGCACTCAAGGCAAGCTTTATTGAGGC-3′. PCR was performed using High Fidelity Platinum Taq and run at 94 °C for 2 min; 45 cycles of 94 °C for 15 s, 55 °C for 30 s and 68 °C for 5 min; and 68 °C for 15 min.

Pseudovirus generation. Selected single genome sequences from outgrowth culture supernatants and plasma were used to generate pseudoviruses that were tested for sensitivity to bNAbs in a TZM-bl neutralization assay. To produce the pseudoviruses, plasmid DNA containing the cytomegalovirus (CMV) promoter was amplified by PCR using forward primer 5′-AGTAATCAATTACGGGGTCATTAGTTCAT-3′ and reverse primer 5′-CATAGGAGATGCCTAAGCCGGTGGAGCTCTGCTTATATAGACCTC-3′. Individual *env* amplicons were amplified using forward primer 5′-CACC GGCTTAGGCATCTCCTATGGCAGGAAGAA-3′ and reverse primer 5′-GTCTCGAGATACTGCTCCCACCC-3′. The CMV promoter amplicon was fused to individual purified *env* amplicons by PCR using forward primer 5′-AGTAATCAATTACGGGGTCATTAGTTCAT-3′ and reverse primer 5′-ACTTTTTGACCACTTGCCACCCAT-3′. Overlapping PCR was carried out using the High Fidelity Platinum Taq (Invitrogen) in a 50-μl reaction consisting of 1 ng purified CMV promoter amplicon, 0.125 μl purified *env* SGA amplicon, 400 nM each forward and reverse primers, 200 μM dNTP mix, 1× Buffer HiFi and 1 μl DNA polymerase mix. PCR was run at 94 °C for 2 min; 25 cycles of 94 °C for 12 s, 55 °C for 30 s and 68 °C for 4 min; and 72 °C for 10 min. Resulting amplicons were analysed by gel electrophoresis, purified by gel extraction, and cotransfected with pSG3Δenv into HEK293T cells to produce pseudoviruses as described previously[47].

Sequence and phylogenetic analysis. Nucleotide alignments of intact *env* sequences were translation-aligned using ClustalW v.2.1[48] under the BLOSUM cost matrix. Sequences with premature stop codons and frameshift mutations that fell in the gp120 surface glycoprotein region were excluded from all analyses. Maximum likelihood phylogenetic trees were then generated from these alignments with PhyML v.3.1[49] using the GTR model with 1,000 bootstraps. For the combined analysis of sequences from all participants, *env* sequences were aligned using MAFFT v.7.309 and clustered using RAxML v.8.2.9[50] under the GTRGAMMA model with 1,000 bootstraps. To analyse changes between reservoir and rebound viruses, *env* sequences were aligned at the amino acid level to a HXB2 reference using ClustalW v.2.1.

Statistical analyses. For sample size considerations, one-sided Clopper-Pearson confidence intervals were calculated for varying number of observed rebounds. A sample size of 15 HIV-1 infected individuals was determined to allow for the rejection of the null hypothesis (rate = 0.85) with 80% power for an effect size equal to or higher than 0.33, if at least 6 out of 15 enrolled participants did not experience viral rebound by week 8 (2 weeks after the last antibody infusions). Pharmacokinetic parameters were estimated by performing a non-compartmental analysis using Phoenix WinNonlin Build 8 (Certara), using all available PK data starting with the time point after the last infusion of 10-1074 from either TZM-bl assay (using the X2088_c9 pseudovirus to determine 10-1074 levels) or ELISA, and compared by a two-tailed unpaired Student's *t*-test. $CD4^+$ T cell counts on day 0 and at the time of viral rebound were compared by two-tailed paired Student's *t*-test. IUPMs determined at week −2 and week 12 were compared using a two-tailed paired Student's *t*-test. Time to rebound in current trial participants (combination therapy with 3BNC117 and 10-1074), participants receiving 3BNC117 monotherapy[9] and participants in previous non-interventional ATI studies conducted by ACTG[10] were plotted using Kaplan–Meier survival curves. Potential correlation between IUPM and time to rebound was analysed by two-tailed Pearson's correlations.

Recombination analysis of *env* sequences. Multiple alignment of nucleotide sequences guided by amino acid translations of *env* sequences was performed by TranslatorX (http://translatorx.co.uk/). Latent and rebound sequences were analysed for the presence of recombination using the 3SEQ recombination algorithm (http://mol.ax/software/3seq/). Sequences that showed statistical evidence of recombination (rejection of the null hypothesis of clonal evolution) in which 'parent' sequences were derived from the latent reservoir and the 'child' sequence was a rebound sequence are represented in a circos plot (http://circos.ca/).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

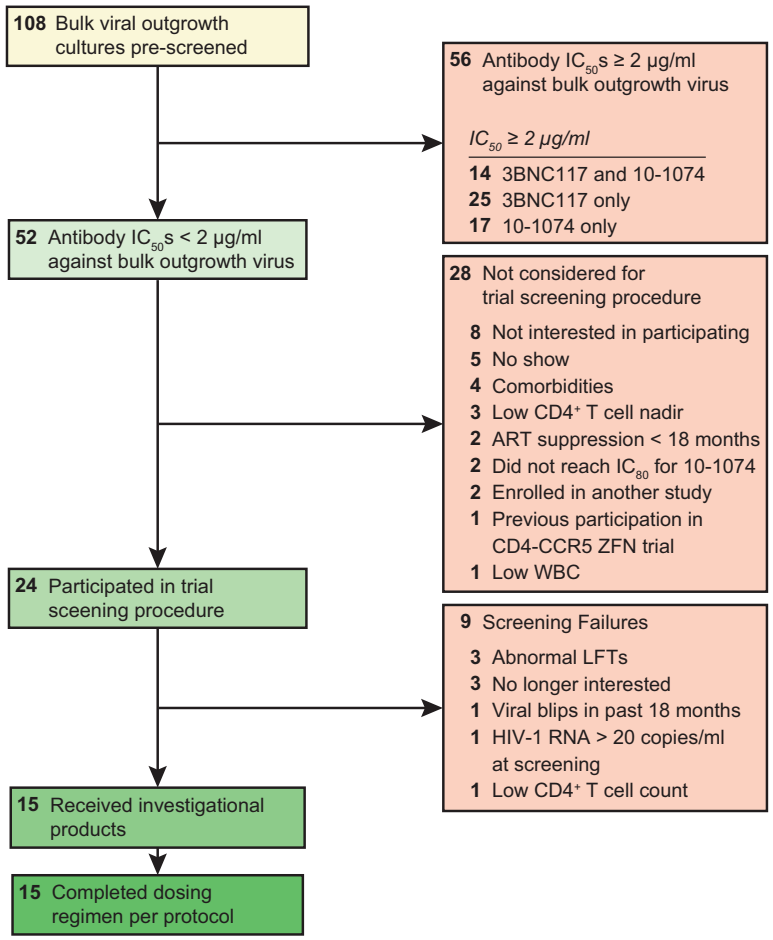
Data availability

The sequences from all isolated viruses are available in GenBank, accession numbers MH575375–MH576416.

46. Salazar-Gonzalez, J. F. et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J. Virol.* **82**, 3952–3970 (2008).
47. Kirchherr, J. L. et al. High throughput functional analysis of HIV-1 *env* genes without cloning. *J. Virol. Methods* **143**, 104–111 (2007).
48. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
49. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
50. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

© 2018 Springer Nature Limited. All rights reserved.

a



b

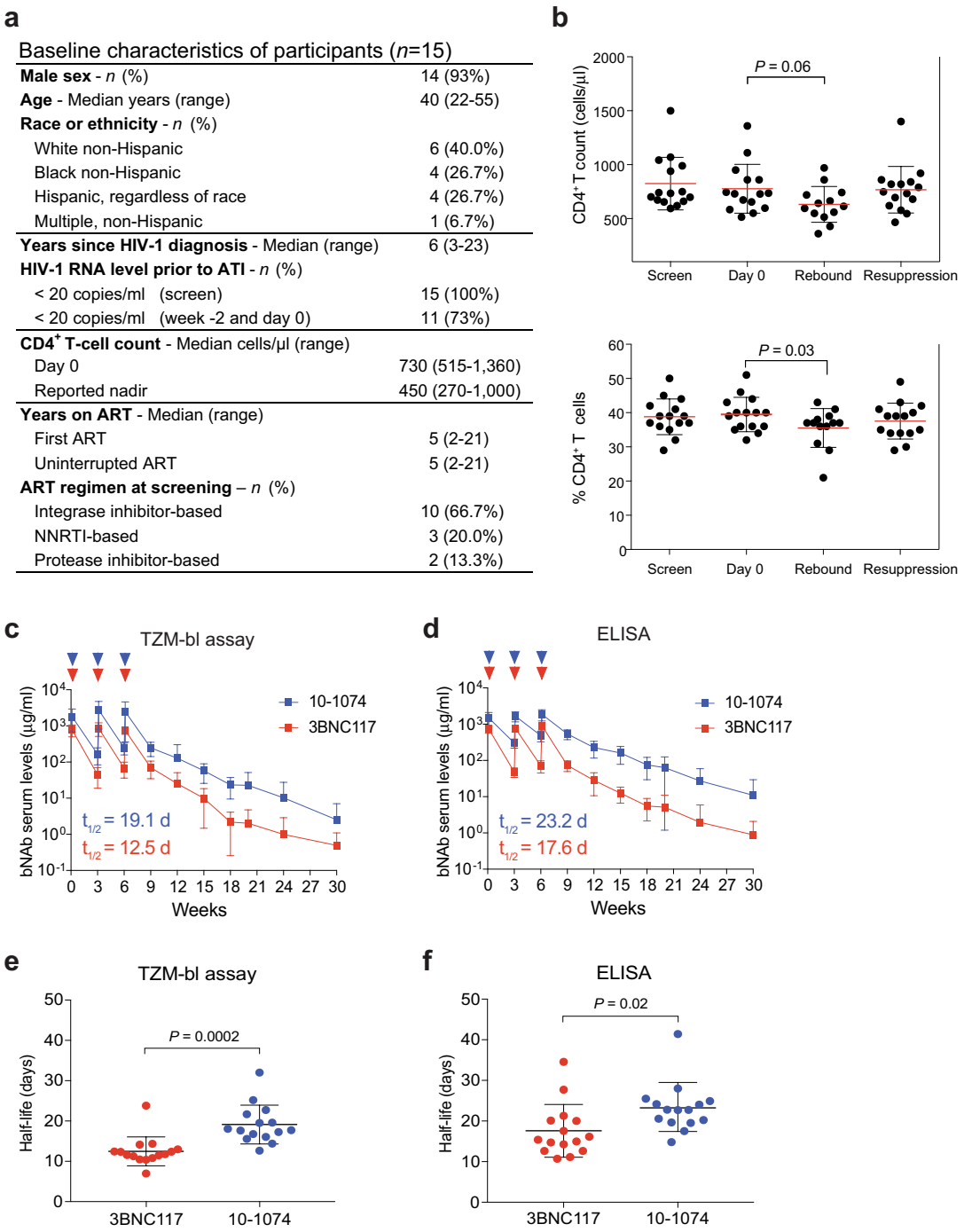
ID	Age	Gender	Race	Years since HIV-1 dx	Years since first ART	Uninterr. ART before ATI (yrs)	ART at Screening*	Switched ART**	Reported CD4 nadir	HLA alleles	Pre-Screen Sensitivity (µg/ml)*** 3BNC117 IC_{50}	3BNC117 IC_{80}	10-1074 IC_{50}	10-1074 IC_{80}	CD4 count (d0)	HIV-1 RNA (cp/ml)# Scr	Wk -2	d0	Weeks to viral rebound
9241	40	M	White/Hisp	6	5	5	EVG/cobi/ TDF/FTC	-	500	n.d.	0.809	2.212	0.090	0.243	515	<20	<20	<20	21
9242	43	M	White/Hisp	3	3	2	EVG/cobi/ TDF/FTC	-	450	n.d.	0.160	0.433	0.144	0.389	654	<20	<20	<20	15
9243	29	M	Amer Indian/Hisp	5	5	5	RPV/TDF/FTC	DTG/TDF/ FTC	350	n.d.	0.641	2.913	0.072	0.241	583	<20	<20 D	<20 D	20
9244	36	M	Amer Indian/not Hisp	9	5	5	EFV/TDF/FTC	DTG/TDF/ FTC	730	n.d.	0.277	0.966	0.025	0.068	1,110	<20	<20	<20	21
9245	22	M	White/Hisp	5	5	5	EVG/cobi/ TAF/FTC	-	360	n.d.	0.417	1.423	0.038	0.089	736	<20	<20	<20	5
9246	30	M	Black	5	5	5	EVG/cobi/ TAF/FTC	-	500	n.d.	0.144	0.387	0.040	0.105	745	<20	<20	<20 D	19
9247	31	M	Black	6	6	6	EVG/cobi/ TAF/FTC	-	600	n.d.	0.556	1.930	0.072	0.326	728	<20	<20	<20	26
9248	52	M	White	11	11	11	DRV/RTV/ TAF/FTC	-	310	n.d.	1.863	9.738	0.676	2.252	730	<20 D	<20 D	58	12
9249	49	M	White	23	20	6	DRV/RTV ABC/3TC	-	426	n.d.	0.562	2.095	0.260	0.983	860	<20 D	<20 D	32	3
9250	55	M	White	7	5	5	EVG/cobi/ TAF/FTC	-	350	n.d.	0.558	3.174	0.447	2.644	550	<20	40	50	6
9251	40	M	Black	6	2	2	EVG/cobi/ TDF/FTC	-	1,000	n.d.	1.200	3.125	0.073	0.153	672	<20	<20	<20	7
9252	51	F	Black	11	11	11	EFV/TDF/FTC	DTG/TDF/ FTC	270	n.d.	0.630	3.074	0.243	0.640	598	<20	<20	<20	22
9253	41	M	White	5	2	2	DTG/TAF/FTC	-	387	n.d.	0.558	2.644	0.020	0.317	950	<20	41	<20 D	5
9254	48	M	White	21	21	21	EVG/cobi/ TAF/FTC	-	590	A1,29 B38,44	0.142	0.386	0.085	0.240	860	<20	<20	<20	>30
9255	30	M	White	5	4	4	EVG/cobi/ TAF/FTC	-	779	A3,25 B18,44	0.324	0.833	0.006	0.015	1,360	<20	<20 D	<20	>30

Extended Data Fig. 1 | Study participant selection and demographics. **a**, Flow diagram indicating the selection of study participants. **b**, Individual participant demographics and baseline clinical characteristics. Grey-shaded rows indicate participants who were found to have detectable viraemia (HIV-1 viral load of >20 copies ml^{-1}) at week −2 or day 0. These participants were not included in the efficacy analyses given the lack of viral suppression at baseline. Amer Indian, American Indian; Hisp, Hispanic. *3TC, lamivudine; ABC, abacavir; cobi, cobicistat; DRV, darunavir; DTG, dolutegravir; EFV, efavirenz; EVG, elvitegravir; FTC, emtricitabine; RPV, rilpivirine; RTV, ritonavir; TAF, tenofovir alafenamide fumarate; TDF, tenofovir disoproxil fumarate. **NNRTI-based regimens were switched four weeks before ART interruption due to longer half-lives of NNRTIs. ***Pre-screening of bulk outgrowth virus obtained from PBMC cultures by TZM-bl assay. #All participants harboured clade B viruses. Viral load <20 D, plasma HIV-1 RNA detected but not quantifiable by clinical assay. d0, day 0; Dx, diagnosis; Scr, screening; Wk −2, week −2.

© 2018 Springer Nature Limited. All rights reserved.

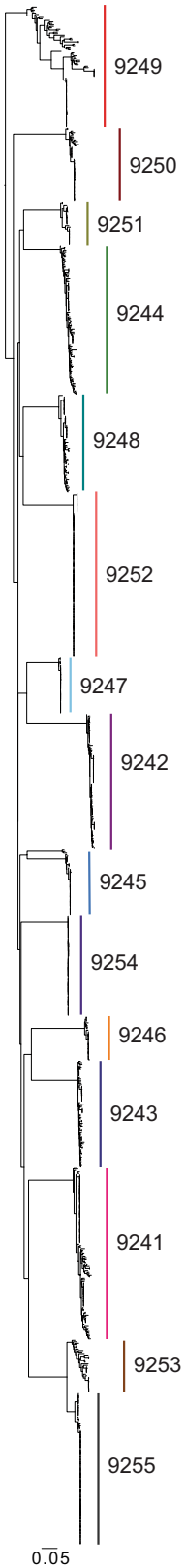
a

Baseline characteristics of participants (n=15)	
Male sex - n (%)	14 (93%)
Age - Median years (range)	40 (22-55)
Race or ethnicity - n (%)	
White non-Hispanic	6 (40.0%)
Black non-Hispanic	4 (26.7%)
Hispanic, regardless of race	4 (26.7%)
Multiple, non-Hispanic	1 (6.7%)
Years since HIV-1 diagnosis - Median (range)	6 (3-23)
HIV-1 RNA level prior to ATI - n (%)	
< 20 copies/ml (screen)	15 (100%)
< 20 copies/ml (week -2 and day 0)	11 (73%)
CD4$^+$ T-cell count - Median cells/µl (range)	
Day 0	730 (515-1,360)
Reported nadir	450 (270-1,000)
Years on ART - Median (range)	
First ART	5 (2-21)
Uninterrupted ART	5 (2-21)
ART regimen at screening – n (%)	
Integrase inhibitor-based	10 (66.7%)
NNRTI-based	3 (20.0%)
Protease inhibitor-based	2 (13.3%)



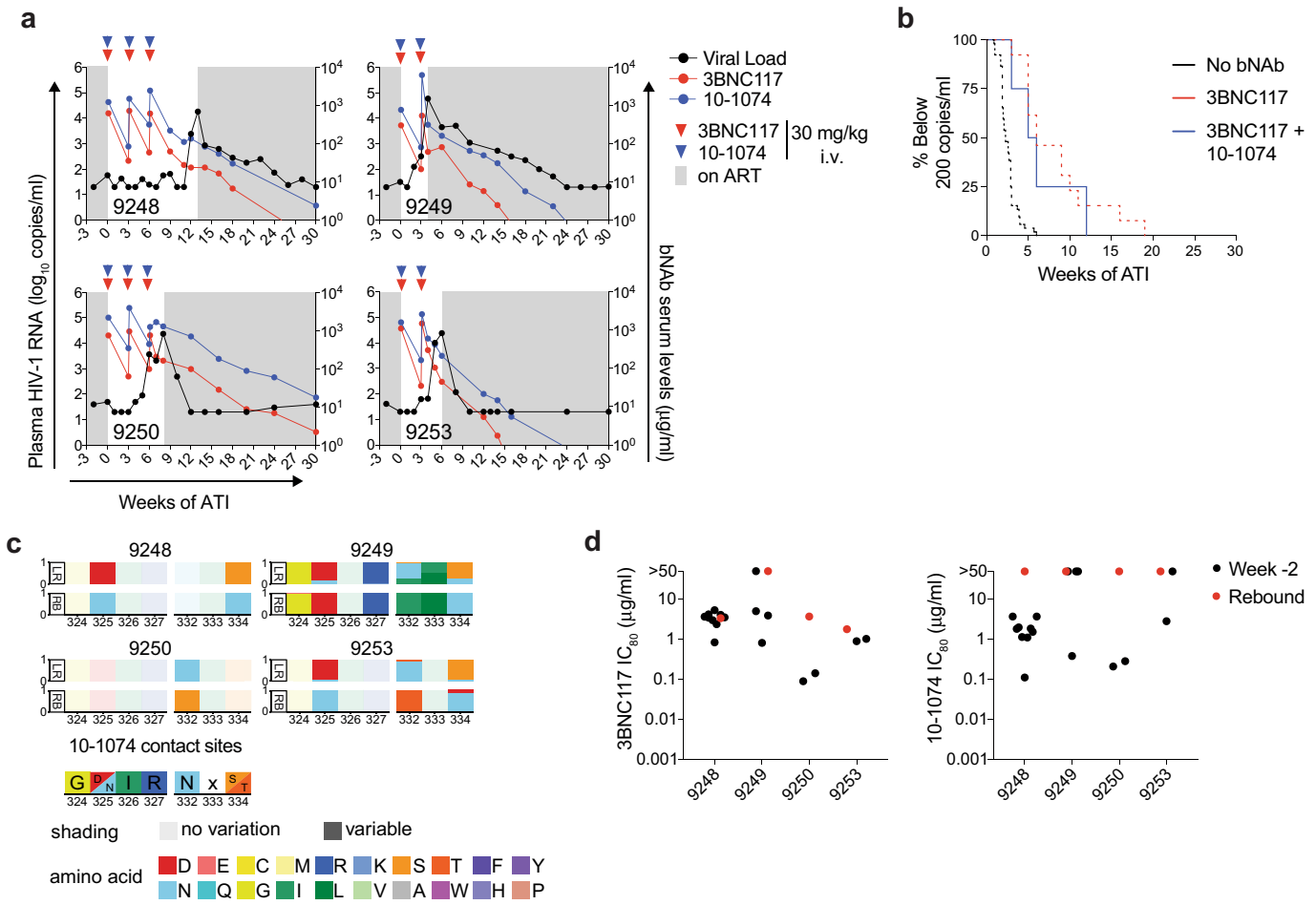
Extended Data Fig. 2 | Demographics, CD4$^+$ T cells during study period in participants and pharmacokinetics of 3BNC117 and 10-1074. **a**, Baseline participant demographics. **b**, Absolute CD4$^+$ T cell counts and percentage of CD4$^+$ T cells among CD3$^+$ T cells at screening (n = 15), day 0 (n = 15), at the time of viral rebound (n = 13) and at the end of the study are shown (n = 15) (see also Supplementary Table 2). The last available time point after resuppression was used as end of the study time point for the participants that reinitiated ART. Red lines indicate mean, error bars indicate standard deviation and individual participants are shown as dots. P values were obtained using a two-tailed paired Student's t-test comparing CD4$^+$ T cell counts between day 0 and the time of viral rebound. **c**, **d**, 3BNC117 (red) and 10-1074 (blue) levels in serum (n = 15) as determined by TZM-bl assay (**c**) and ELISA (**d**). In cases in which participants only received 2 infusions due to early viral rebound (9245, 9249 and 9253), only antibody concentrations up to the second infusion were included. Half-life of each bNAb is indicated in days. Curves indicate mean serum antibody concentrations and error bars represent standard deviation. Red and blue triangles indicate 3BNC117 and 10-1074 infusions, respectively. **c**, In the TZM-bl assay, lower limits of quantification were 0.46 µg ml^{-1} and 0.10 µg ml^{-1} for 3BNC117 and 10-1074, respectively. **d**, In the ELISA, lower limits of detection were 0.78 µg ml^{-1} and 0.41 µg ml^{-1}, respectively. **e**, **f**, Half-lives of both antibodies as measured by TZM-bl assay (**e**) and ELISA (**f**). Each dot represents a single participant. The half-lives of both antibodies from the 15 participants enrolled in the study are represented. Black lines indicate the mean value and standard deviation (n = 15). P values were obtained using a two-tailed unpaired Student's t-test comparing the two antibodies.

© 2018 Springer Nature Limited. All rights reserved.



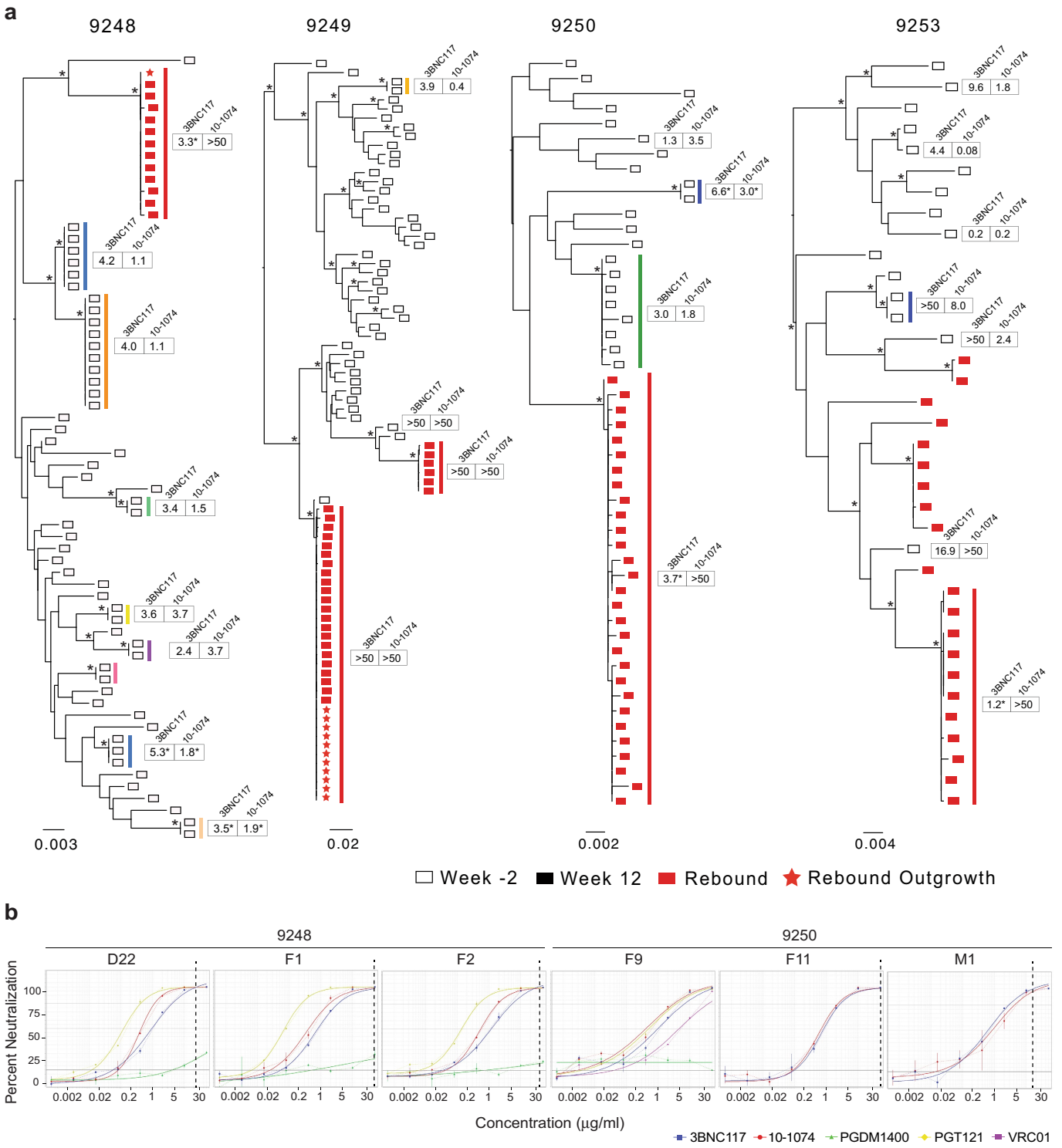
Extended Data Fig. 3 | Phylogenetic tree of viruses from all enrolled participants. Maximum likelihood phylogenetic trees of full-length *env* sequences containing all sequences obtained from Q^2VOA cultures and rebound viruses from SGA or rebound outgrowth of the 15 participants enrolled in the study. Participants are indicated by individual colours.

© 2018 Springer Nature Limited. All rights reserved.



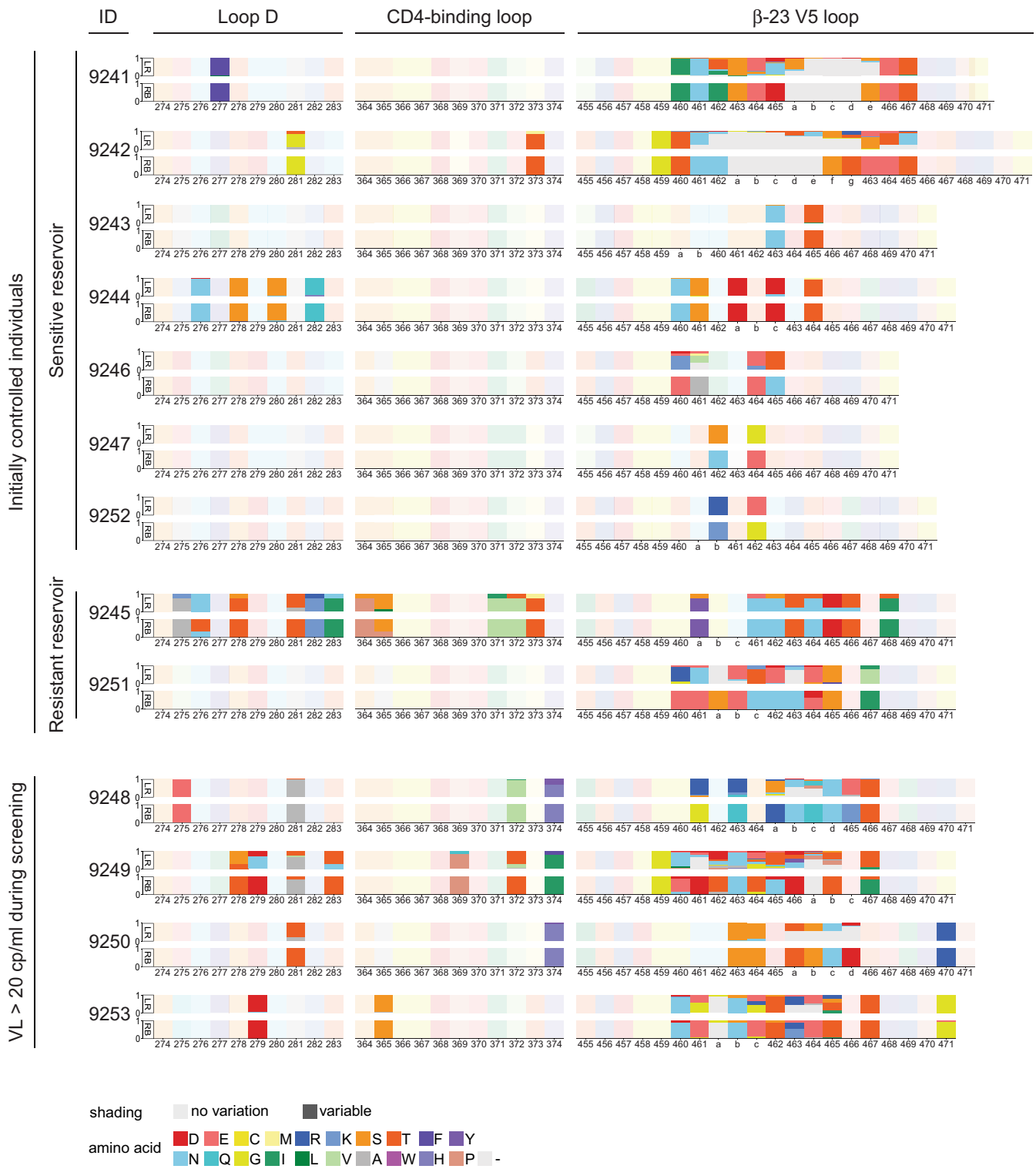
Extended Data Fig. 4 | Viral rebound, amino acid variants at 10-1074 contact sites and sensitivities of latent and rebound viruses in the participants with detectable viraemia (>20 copies per ml) two weeks before or at the start of ATI. a, Plasma HIV-1 RNA levels (black; left *y* axis) and bNAb serum concentrations (3BNC117, red; 10-1074, blue; right *y* axis). Red and blue triangles indicate 3BNC117 and 10-1074 infusions, respectively. Serum antibody concentrations were determined by TZM-bl assay. Grey-shaded areas indicate time on ART. Lower limit of detection of HIV-1 RNA was 20 copies per ml. **b**, Kaplan–Meier plots summarizing time to viral rebound. The *y* axis shows the percentage of participants that maintained viral suppression. The *x* axis shows the time in weeks after the start of ATI. Participants receiving the combination of 3BNC117 and 10-1074 are indicated by the blue line ($n = 4$). The dotted red line indicates a cohort of individuals receiving 3BNC117 alone during ATI[9] ($n = 13$) and the dotted black line indicates a cohort of participants who underwent ATI without any intervention[10] ($n = 52$). **c**, Colour charts show Env contact sites of 10-1074 at the G(D/N)IR motif (positions 324–327, according to HXB2 numbering) and the glycan at the potential *N*-linked glycosylation site at position 332 (NxS/T motif at positions 332–334). LR, latent reservoir viruses isolated by Q^2VOA (week −2); RB, rebound viruses isolated by SGA (plasma) or viral outgrowth (PBMCs). Each amino acid is represented by a colour and the frequency of each amino acid is indicated by the height of the rectangle. Shaded rectangles indicate the lack of variation between latent reservoir and rebound viruses at the indicated position. Full-colour rectangles represent amino acid residues with changes in distribution between reservoir and rebound viruses. **d**, Dot plots showing the IC_{80} (μg ml^{-1}) of 3BNC117 (left) and 10-1074 (right) against latent and rebound viruses determined by TZM-bl neutralization assay. Q^2VOA-derived latent viruses from week −2 are shown as black circles. For outgrowth culture-derived rebound viruses, the highest IC_{80} determined is shown as red circle. For 9250 and 9253, no viruses could be obtained from rebound outgrowth cultures and pseudoviruses were made from *env* sequences of the latent reservoir (Q^2VOA) and rebound viruses (plasma SGA). Note that 9249 and 9253 had pre-existing resistant viruses in the reservoir ($IC_{50} > 2$ μg ml^{-1}). 9248 and 9250 had pre-existing viruses that failed to reach an IC_{100} when tested up to 50 μg ml^{-1} for 3BNC117 (Extended Data Fig. 5). Rebound viruses of all four participants had an IC_{80} or IC_{100} of >50 μg ml^{-1} for both 3BN117 and 10-1074.

© 2018 Springer Nature Limited. All rights reserved.



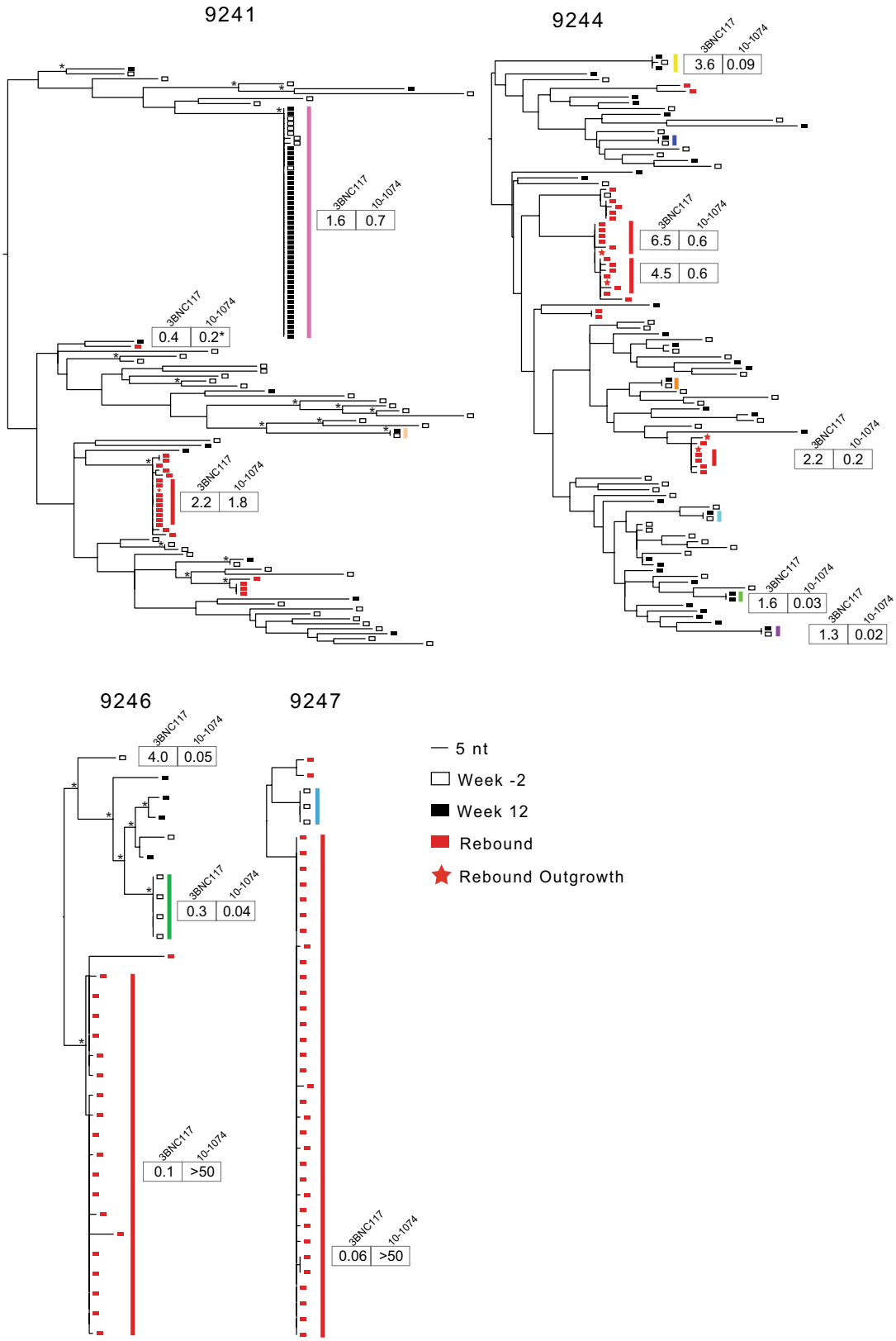
Extended Data Fig. 5 | Phylogenetic *env* trees and TZM-bl neutralization curves for individuals with viral blips. a, Circulating reservoir and viral rebound in study participants with detectable viraemia at week −2 or day 0. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from week −2 Q^2VOA cultures, rebound plasma SGA and rebound outgrowth from the four participants with viral blips. Open black rectangles indicate Q^2VOA-derived viruses from week −2. Viruses obtained at the time of rebound are indicated by red rectangles (plasma SGA) and red stars (rebound PBMC outgrowth cultures), respectively. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥70%). Clones are denoted by coloured lines. Boxes indicate IC_{80} values (μg ml^{-1}) of 3BNC117 and 10-1074 against individual clones, with asterisks indicating $IC_{100} > 50$ μg ml^{-1}. **b**, Latent reservoir virus TZM-bl neutralization curves for two participants that had a viral load of >20 copies per ml at day 0 (9248 and 9250). Curves show neutralization titres by 3BNC117 (blue), 10-1074 (red) and other bNAbs, when available, for week −2 Q^2VOA-derived viruses present in the circulating reservoir. Three representative viruses from 9248 (left) and 9250 (right) are shown. Although these viruses had low 3BNC117 and 10-1074 IC_{50} or IC_{80} titres, the IC_{100} (black dotted line) is reached only at a high concentration or not reached at all. The neutralization titre was measured by TZM-bl neutralization assay using a five-parameter curve fit method.

© 2018 Springer Nature Limited. All rights reserved.



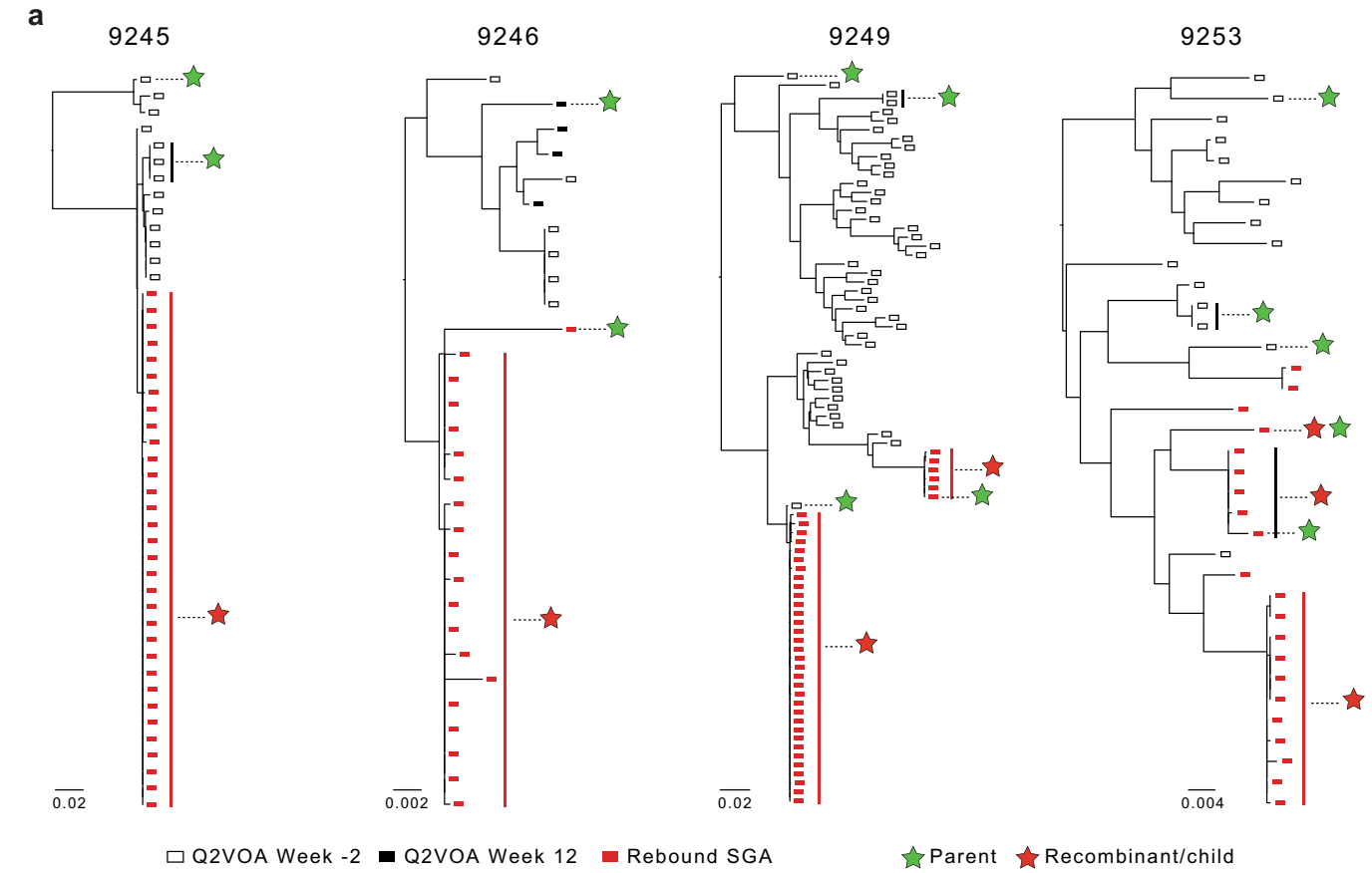
Extended Data Fig. 6 | Amino acid variants at 3BNC117 contact sites of reactivated latent and rebound viruses. Colour charts show 3BNC117 contact sites in Env according to HXB2 numbering. Diagram shows the 13 participants that experienced viral rebound before week 30. LR, latent reservoir viruses isolated by Q^2VOA (on weeks −2 and 12 when available); RB, rebound viruses isolated by SGA (plasma) and viral outgrowth (PBMCs). Each amino acid is represented by a colour and the frequency of each amino acid is indicated by the height of the rectangle. Shaded rectangles indicate the lack of variation and full-colour rectangles represent amino acid residues with changes in the distribution between the reservoir and rebound.

© 2018 Springer Nature Limited. All rights reserved.



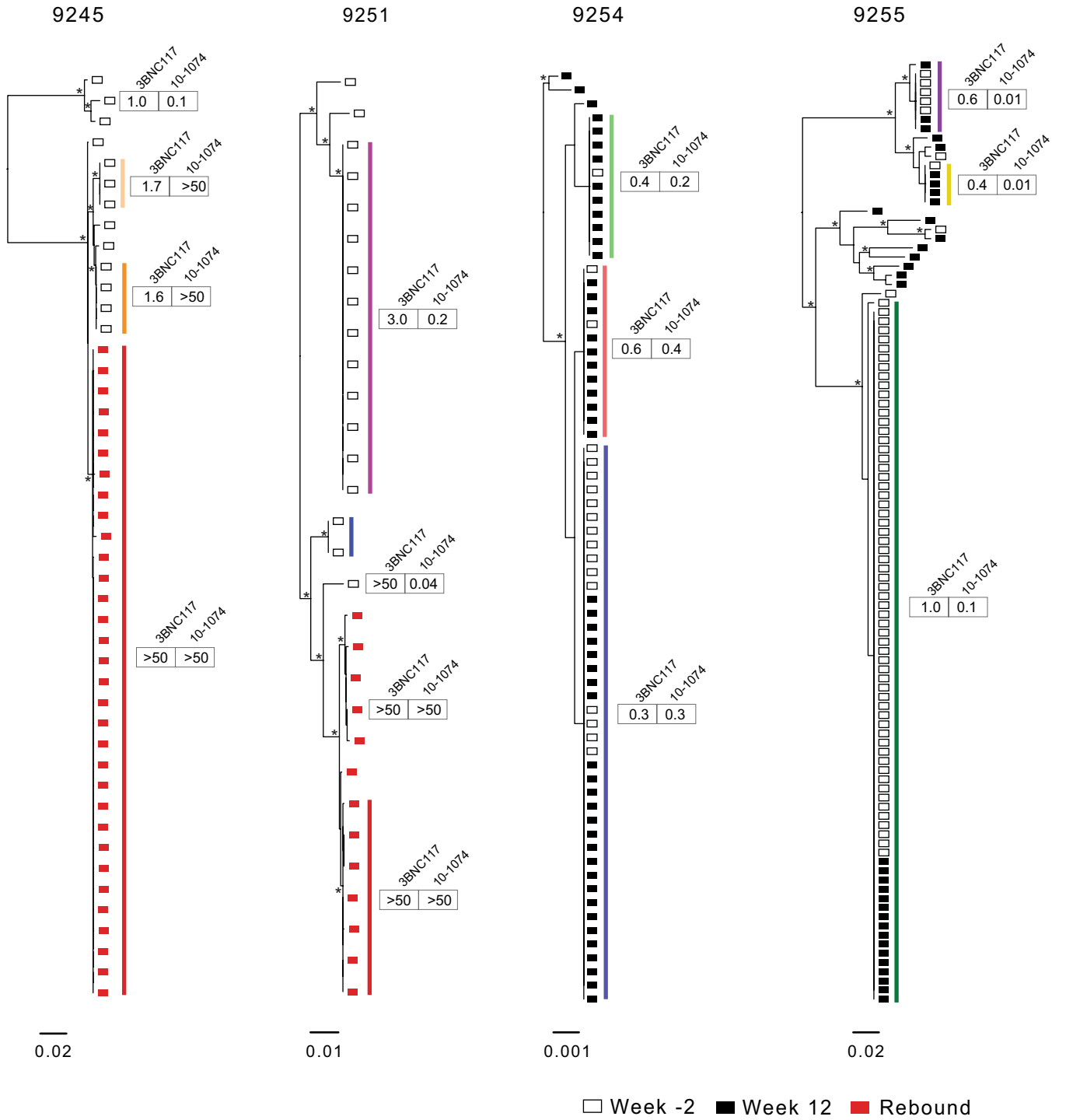
Extended Data Fig. 7 | Comparison of the circulating latent reservoir and rebound viruses. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q^2VOA, rebound plasma SGA and rebound PBMC outgrowth cultures from participants 9241, 9244, 9246 and 9247, who rebounded before week 30. Open and closed black rectangles indicate Q^2VOA-derived viruses from week −2 and week 12, respectively. Viruses obtained at the time of rebound are indicated by red rectangles (plasma SGA) and red stars (rebound PBMC outgrowth cultures). Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥70%). Clones are denoted by coloured lines mirroring the colours of slices in Extended Data Fig. 10a. Boxes indicate IC_{80} values (μg ml^{-1}) of 3BNC117 and 10-1074 against representative viruses throughout the phylogenetic tree and clones, when possible (Supplementary Table 4). Asterisks in boxes indicate $IC_{100} > 50$ μg ml^{-1}.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Recombination events in rebound viruses. **a**, Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q^2VOA cultures and rebound SGA in the four participants for whom rebound viruses showed recombination events. Open and closed black rectangles indicate Q^2VOA-derived viruses from week −2 and week 12, respectively. Rebound plasma SGA- or outgrowth-derived viruses are indicated by closed red rectangles. Green stars represent parent sequences that underwent recombination to produce the child sequences (red stars). **b**, Circos plots indicating the relationship between the parent sequences and the recombinants. Open and closed black rectangles indicate Q^2VOA-derived sequences from week −2 and week 12, respectively. Rebound virus sequences are indicated by red rectangles. The thickness of the black outer bars represents the number of sequences obtained from that particular clone.

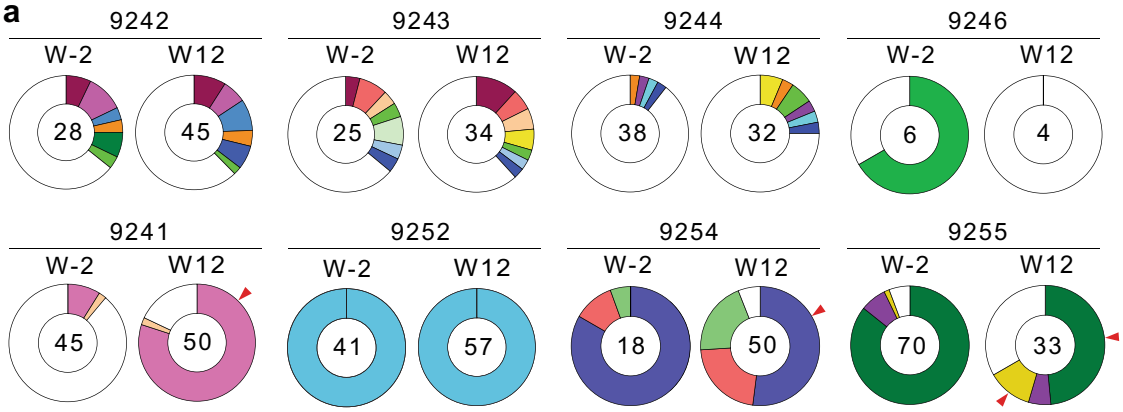
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Phylogenetic trees of participants 9245, 9251, 9254 and 9255. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q^2VOA cultures and rebound plasma SGA and rebound outgrowth from the two participants (9245 and 9251) with pre-existing resistance to one of the two antibodies and the two sensitive participants (9254 and 9255) who maintained viral suppression for >30 weeks (end of the study). Open and closed black rectangles indicate Q^2VOA-derived viruses from week −2 and week 12, respectively. Rebound plasma SGA viruses are indicated by closed red rectangles. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥ 70%). Clones are denoted by coloured lines beside the phylogenetic tree. Numbers correspond to 3BNC117 and 10-1074 IC_{80} neutralization titres.

© 2018 Springer Nature Limited. All rights reserved.

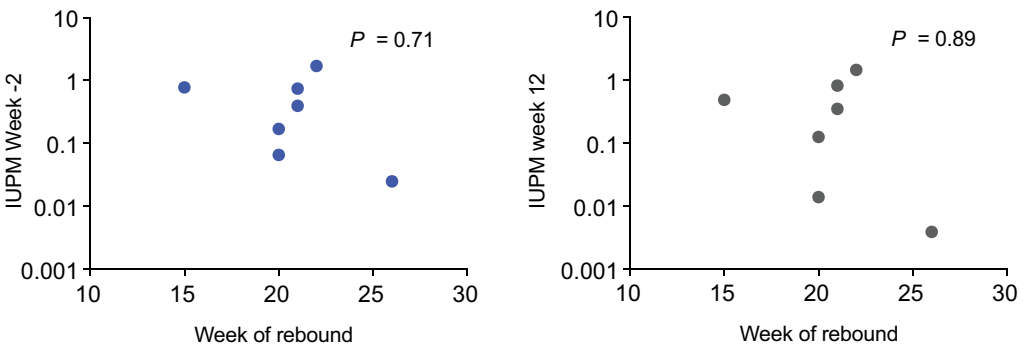
a



b

Study ID	Total *env* seqs by Q^2VOA	Clonal *env* sequences		IUPM	
	(no.)	(no.)	(%)	week -2	week 12
9242	73	27	37.0	0.781	0.493
9243	59	22	37.3	0.170	0.126
9244	70	12	17.1	0.397	0.354
9246	10	4	40.0	0.066	0.014
9247	3	3	100.0	0.025	0.004
9241	95	46	48.4	0.743	0.828
9252	98	98	100.0	1.709	1.470
9254	68	65	95.6	N/A	0.680
9255	103	88	85.4	1.890	1.400
Total	**579**	**365**	**63.0**		

c



Extended Data Fig. 10 | Clonal distribution of the circulating latent reservoir and IUPM changes. a, Pie charts depicting the distribution of Q^2VOA-derived *env* sequences obtained at weeks −2 (W−2) and week 12 (W12). Number in the inner circle indicates the total number of analysed *env* sequences. White represents sequences isolated only once across both time points and coloured slices represent identical sequences that appear more than once (clones). The size of each pie slice is proportional to the size of the clone. Red arrows denote clones that significantly change in size ($P \leq 0.05$ (two-sided Fisher's exact test)) between the two time points. **b**, Summary of clonal *env* sequences and IUPM in the nine individuals with an antibody-sensitive reservoir. **c**, IUPM versus time of viral rebound in the antibody-sensitive individuals ($n = 7$) who rebounded within the study observation period (30 weeks). *P* values were obtained using a two-tailed Pearson correlation test comparing the two variables.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0509-0

De novo design of a fluorescence-activating β-barrel

Jiayi Dou[1,2,11], Anastassia A. Vorobieva[1,2,11], William Sheffler[1,2], Lindsey A. Doyle[3], Hahnbeom Park[1,2], Matthew J. Bick[1,2], Binchen Mao[1,7], Glenna W. Foight[4], Min Yen Lee[4], Lauren A. Gagnon[4], Lauren Carter[1,2], Banumathi Sankaran[5], Sergey Ovchinnikov[1,2,8], Enrique Marcos[1,2,9], Po-Ssu Huang[1,2,10], Joshua C. Vaughan[4], Barry L. Stoddard[3] & David Baker[1,2,6]*

The regular arrangements of β-strands around a central axis in β-barrels and of α-helices in coiled coils contrast with the irregular tertiary structures of most globular proteins, and have fascinated structural biologists since they were first discovered. Simple parametric models have been used to design a wide range of α-helical coiled-coil structures, but to date there has been no success with β-barrels. Here we show that accurate de novo design of β-barrels requires considerable symmetry-breaking to achieve continuous hydrogen-bond connectivity and eliminate backbone strain. We then build ensembles of β-barrel backbone models with cavity shapes that match the fluorogenic compound DFHBI, and use a hierarchical grid-based search method to simultaneously optimize the rigid-body placement of DFHBI in these cavities and the identities of the surrounding amino acids to achieve high shape and chemical complementarity. The designs have high structural accuracy and bind and fluorescently activate DFHBI in vitro and in *Escherichia coli*, yeast and mammalian cells. This de novo design of small-molecule binding activity, using backbones custom-built to bind the ligand, should enable the design of increasingly sophisticated ligand-binding proteins, sensors and catalysts that are not limited by the backbone geometries available in known protein structures.

There have been considerable recent advances in designing protein folds from scratch[1,2], as well as redesigning already existing native scaffolds to bind small molecules[3–5], but two outstanding unsolved challenges remain. The first is the de novo design of all-β proteins, which is complicated by the tendency of β-strands and sheets to associate intermolecularly to form amyloid-like structures if their register is not perfectly controlled[6]. The second is the design of proteins customized to bind small molecules of interest, which requires precise control over backbone and side chain geometry[5], as well as the balancing of the often opposing requirements of protein folding and function[7]. Success in developing such methods would reduce the long-standing dependency on natural proteins by enabling protein engineers to craft new proteins optimized to bind chosen small-molecule targets, and lay a foundation for de novo design of proteins customized to catalyse specific chemical reactions.

Principles for designing β-barrels

β-barrels are single β-sheets that twist to form a closed structure in which the first strand is hydrogen-bonded to the last[8]. Anti-parallel β-barrels are excellent scaffolds for ligand binding, as the base of the barrel can accommodate a hydrophobic core to provide overall stability, and the top of the barrel can provide a recessed cavity for ligand binding[9], often flanked by loops that can contribute further binding affinity and selectivity[10]. However, all β-sheet topologies are notoriously difficult to design from scratch—to our knowledge, there has been no reported success to date—although several descriptive parametric models of β-barrels have been proposed[11–13]. We first set out to address this challenge by parametrically generating regular arrangements of eight anti-parallel β-strands using the equations for an elliptic hyperboloid of revolution (adapted from previously published work[14], Fig. 1a). β-barrels are characterized by their shear number (S)—the total shift in strand registry between the first and last strands—which determines the hydrophobic packing arrangement and the diameter of the barrel[15,16] (Supplementary Methods). We selected a shear number of $S = 10$ because it is difficult to achieve good core packing for $S = 8$ (the barrel has a smaller diameter and the C_α–C_β vectors point directly at each other), and $S = 12$ results in a cavity that is too large to fill with side chains (Extended Data Fig. 1b–d). We generated ensembles of hyperboloids by sampling the elliptical parameters and the tilt of the generating lines with respect to the central axis around ideal values computed for $S = 10$, and then placed C_α atoms on the hyperboloid surface (Fig. 1a, Supplementary Methods). As found in earlier simulation work[17], backbones generated with constant angles between strands could not achieve perfectly regular hydrogen bonding. To resolve this problem, we introduced force-field guided variation in local twist by gradient-based minimization. We selected the backbones with the most extensive inter-strand hydrogen bonding, connected the strands with short loops and carried out combinatorial sequence optimization to obtain low-energy sequences (Extended Data Fig. 1a). Synthetic genes encoding 41 such designs were produced and the proteins were expressed in *E. coli*. Almost all were found to be insoluble or oligomeric; none of this first set of 41 designs were monomeric with an all-β circular dichroism spectrum (Supplementary Table 2).

In considering the possible reasons for the failure of the initial designs, we noted that many of the backbone hydrogen-bond interactions on the top and bottom of the barrels were distorted or broken (Extended Data Fig. 1e, f). To investigate the origins of this distortion, we experimented with three alternative approaches to generating

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Division of Basic Science, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [4]Department of Chemistry, University of Washington, Seattle, WA, USA. [5]Molecular Biophysics and Integrated Bioimaging, Berkeley Center for Structural Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [7]Present address: Crown Bioscience, Taicang, China. [8]Present address: John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA. [9]Present address: Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [10]Present address: Department of Bioengineering, Stanford University, Stanford, CA, USA. [11]These authors contributed equally: Jiayi Dou, Anastassia A. Vorobieva. *e-mail: dabaker@uw.edu

© 2018 Springer Nature Limited. All rights reserved.

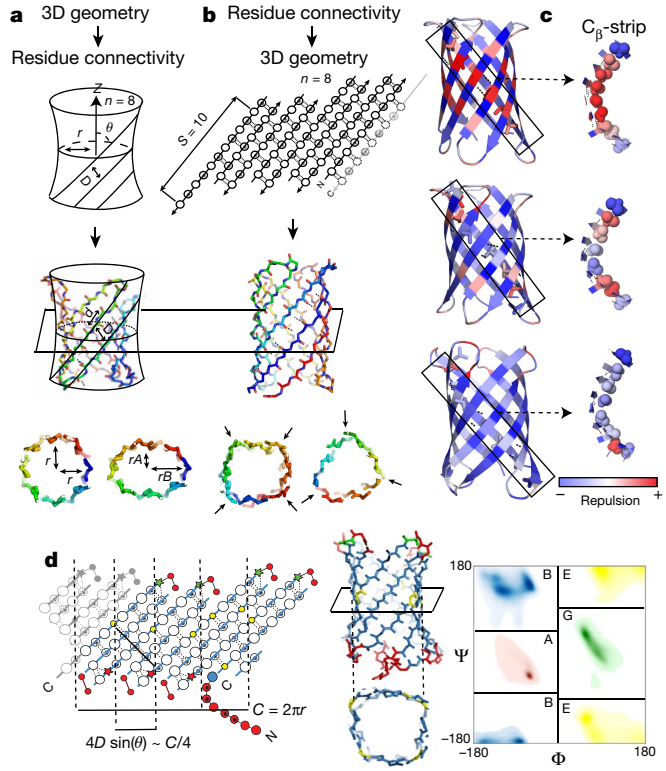


Fig. 1 | Principles for designing β-barrels. a, **b**, Two methods for β-barrel backbone generation. **a**, Parametric generation of 3D backbones on the basis of the hyperboloid model. The cross-section of the barrel is controlled on the global level with parameters (r, r_A and r_B). **b**, Specification of residue connectivity in a 2D map and assembly of 3D backbones with Rosetta. The cross-section geometry is controlled on the local level with torsion angle bins specified for each residue. **c**, Incorporation of glycine kinks and β-bulges reduces Lennard–Jones repulsive interactions in β-barrels. Full backbones are shown on the left and one C_β-strip is shown on the right. Top, no β-bulge, no glycine kink; middle, one glycine kink in the middle of each C_β-strip, no β-bulge; bottom, one glycine kink in the middle of each C_β-strip, and β-bulges placed near the β-turns. **d**, Blueprint used to generate a β-barrel of type (strand number $n = 8$; shear number $S = 10$) with a square cross-section suitable for ligand binding. The values of the barrel radius (r) and tilt of the strands (θ) used to place glycines are determined by the choice of n and S. The residues in the 2D blueprint (left) and the 3D structure (middle) are coloured by backbone torsion bins (right, ABEGO types nomenclature from Rosetta). Shaded and open circles represent residues facing the barrel interior and exterior, respectively. Glycine positions are shown as yellow circles and β-bulges as stars. The 'corners' in the β-sheet resulting from the presence of glycine kinks are shown as vertical dashed lines. C, barrel circumference; D, distance between strands.

uniform β-barrel backbones lacking loops and with valine at every position as a place-holder (Supplementary Methods). In all cases, we observed breaking of hydrogen-bond interactions after structure minimization with the Rosetta relaxation protocol (Extended Data Fig. 2a), suggesting that there is strain inherent to the closing of the curved β-sheet on itself. To identify the origin of this strain, we repeated the relaxation after imposing strong constraints on the hydrogen-bond interactions to prevent them from breaking. The strain manifested in two places. First, steric clashes build up along strips of side chains in the directions of the hydrogen bonds, perpendicularly to the direction of the β-strands ('C_β-strips', Fig. 1c). Second, a number of residues acquired unfavourable left-handed twist (Extended Data Fig. 2b, c; the chirality of the peptide backbone favours right-handed twist). To reduce the strain arising from steric clashes between C_β atoms, and from the local left-handed twist, we replaced the central valine residue of each C_β-strip with glycine residues (which are normally disfavoured in β-sheets[18]). The achiral glycine can have a left-hand twist without disrupting the β-sheet hydrogen-bond pattern[15,19] and lacks a C_β atom, reducing the steric clashes within C_β-strips (Fig. 1c, middle). The backbones of most of these glycine residues shifted to the positive Φ torsion bin after minimization, forming torsional irregularities in the β-sheet ('glycine kinks'[15], Extended Data Fig. 2d–e).

On the basis of these observations, we hypothesized that large local deviations in the ideal β-strand twist are necessary to maintain continuous hydrogen-bond interactions between strands in a closed β-barrel, and hence that a parametric approach assuming uniform geometry was not well-suited to building such structures. Therefore, we chose to build β-barrel backbones starting from a 2D map specifying the peptide bonds, the backbone torsion angle bins[20] and the backbone hydrogen bonds (Fig. 1b). In contrast to parametric backbone design, which may be viewed as a '3D-to-2D' approach as a 3D surface is generated and then populated with residues, this alternative strategy proceeds from 2D to 3D and can readily incorporate local torsional deviation (Fig 1a). We generated 3D protein backbones using Rosetta Monte Carlo structure generation calculations starting from an extended peptide chain[21], guided by torsional and distance constraints from the 2D map.

We found that we could control the volume and the 3D shape of the β-barrel cavity by altering the placement of glycine kinks in the 2D map. Such kinks increase local β-sheet curvature, forming corners in an otherwise roughly circular cross-section (Extended Data Fig. 2f, g). We chose to design a square barrel shape and created four corners in the β-sheet by placing five glycine kinks to eliminate strain in the five C_β-strips, and one glycine kink to adjust the curvature of the longest hairpin (Fig. 1d, Extended Data Fig. 3a, Supplementary Methods). With this choice, the resulting 3D backbones have a large interior volume suitable for a ligand-binding cavity. When such backbones were built with canonical type I′ β-turns connecting each β-hairpin, we observed steric strain at the extremities of the C_β-strips (Fig. 1c, centre) and disruption of hydrogen-bond interactions after structure relaxation (Extended Data Fig. 3e). This probably arises because the considerable curvature at the glycine kinks requires that the β-hairpins paired with it (dashed vertical line in Extended Data Fig. 3b) have greater right-handed twist than can be achieved with canonical β-turns. We reasoned that accentuated right-handed twist could be achieved by incorporating β-bulges—disruptions of the regular hydrogen-bonding pattern of a β-sheet[2,22,23]. Indeed, we found that strategic placement of β-bulges on the bottom of the barrel (defined as the side of the N and C termini) and bulge-containing β-turns[22] on the top of the barrel eliminated steric strain and stabilized the hydrogen bonds between the β-strand residues flanking the turns (Fig. 1c, bottom, Extended Data Fig. 3e, f). To tie together the bottom of the barrel, we introduced a 'tryptophan corner'[24,25] by placing a short 3–10 helix, a glycine kink and a Trp at the beginning of the barrel, and an interacting Arg at the C terminus (Extended Data Fig. 3g–j).

Five hundred backbones were generated from the 2D map incorporating the above features, and Rosetta flexible-backbone sequence design calculations were carried out to identify low-energy sequences for each backbone. Four designs with low energy and backbone hydrogen bonding throughout the barrel were selected for experimental characterization (Extended Data Fig. 4a). The sequences of these designs are not related to those of known native proteins (BLAST E values greater than 0.1), and fold into the designed structure in silico (Fig. 2a). Synthetic genes encoding the designs were expressed in *E. coli*. Three of the designs were expressed in the soluble fraction and purified; two had characteristic β-sheet far-ultraviolet circular dichroism (CD) signal (Fig. 2, Extended Data Fig. 4b). Size-exclusion chromatography (SEC) coupled with multi-angle light scattering (MALS) showed that one was a stable monomer (BB1) and the other (BB2) a soluble tetramer (Extended Data Fig. 4c).

BB1 exhibited a strong near-ultraviolet CD signature, which suggests an organized tertiary structure (Fig. 2d). The design was stable at 95 °C, and cooperatively unfolded in guanidine-denaturation experiments (Fig. 2e). The crystal structure of BB1 solved at 1.6 Å resolution was

© 2018 Springer Nature Limited. All rights reserved.

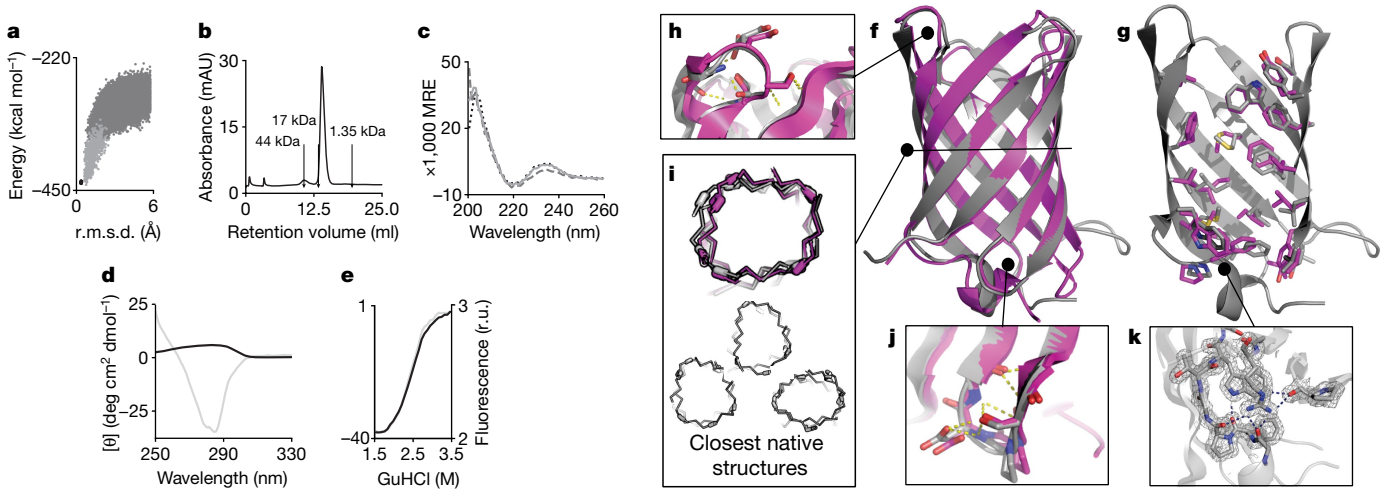


Fig. 2 | Folding, stability and structure of design BB1. a, In silico folding energy landscape. Each grey dot indicates the result of an independent ab initio folding calculation; black dots show results of refinement trajectories starting from design model and light grey dots from the lowest-energy ab initio models. **b**, Size-exclusion chromatogram of the purified monomer (14 kDa). **c**, Far-ultraviolet CD spectra at 25 °C (grey line), 95 °C (black dashed line) and cooled back to 25 °C (black dotted line). **d**, Near-ultraviolet CD spectra in Tris buffer (grey line) and 7 M GuHCl (black line). **e**, Cooperative unfolding in GuHCl monitored by near-ultraviolet CD signal at 285 nm (grey line) and tryptophan fluorescence (black line). **f–j**, Superpositions of the crystal structure (grey) and the design model (pink): overall backbone superposition (**f**); section along the β-barrel axis showing the rotameric states of core residues (**g**); one of the top loop with a G1 β-bulge (**h**); and equatorial cross-section of the β-barrel, showing the geometry of the interior volume (**i**). The glycine kinks are shown as sticks. The bottom of the panel shows the cross-sections of the three closest native β-barrel structures on the basis of TM-score[33] (RCSB Protein Data Bank (PDB) IDs: 1JMX (0.77); 4IL6 (chain O) (0.73); 1PBY (0.71)). **j**, One of the bottom loops with a classic β-bulge. **k**, Crystal structure and $2mF_o - DF_c$ electron density of the tryptophan corner, contoured at 1.5σ.

very close to the design model (1.4 Å backbone root mean squared deviation (r.m.s.d.) over 99 of 109 residues; Extended Data Fig. 4d–f). Essentially all of the key features of the design model are found in the crystal structure (Fig. 2f–k). The barrel cross-section in the crystal structure is very similar to that of the design model, with an overall square shape with corners at the glycine kinks. Natural β-barrel crystal structures do not have this shape; the cross-sections of the closest structure matches in the PDB are shown in Fig. 2i. All seven designed β-turns and β-bulges are correctly recapitulated in the crystal structure (Fig. 2h, j), along with the 3–10 helix and tryptophan corner (Fig. 2k).

Design of small-molecule-binding β-barrels

Having determined principles for de novo design of β-barrels, we next sought to design functional β-barrels with binding sites tailored for a small molecule of interest. We chose DFHBI (Fig. 3a, left, green), a derivative of the intrinsic chromophore of GFP, to test the computational design methods. Owing to its internal torsional flexibility in solution, DFHBI does not fluoresce unless it is constrained in the planar *Z* conformation[26,27]. We sought to design protein sequences that fold into a stable β-barrel structure with a recessed cavity lined with side chains to constrain DFHBI in its fluorescent planar conformation. We chose to take a three step approach: (1) de novo construction of β-barrel backbones, (2) placement of DFHBI in a dedicated pocket, and (3) energy-based sequence design. For the first step, we stochastically generated 200 β-barrel backbones on the basis of the 2D map described above (Extended Data Fig. 5b–d).

The placement of the ligand in the binding pocket requires sampling of both the rigid-body degrees of freedom of the ligand, and the sequence identities of the surrounding amino acids that form the binding site. Because of the dual challenges associated with optimization of structure and sequence simultaneously, most approaches to designing ligand-binding sites to date have separated sampling into two steps: rigid-body placement of the target ligand in the protein-binding pocket and then design of the surrounding sequence[4,5,28]. This two-step approach has the limitation that the optimal rigid-body placement cannot be determined independently of knowledge of the possible interactions with the surrounding amino acids. The RosettaMatch method[29] can identify rigid-body and interacting-residue placements simultaneously, but is limited to a small number of pre-defined ligand-interacting residues[3]. We addressed these challenges with a new 'rotamer interaction field (RIF)' docking method that simultaneously samples over rigid-body and sequence degrees of freedom. RIF docking first generates an ensemble of billions of discrete amino acid side chains that make hydrogen-bonding and non-polar hydrophobic interactions with the target ligand (Fig. 3a, right). Then, scaffolds are docked into this pre-generated interacting ensemble using a grid-based hierarchical search algorithm (Extended Data Fig. 5a). We used RIF docking to place DFHBI into the upper half of the β-barrel scaffolds, resulting in 2,102 different ligand–scaffold pairs with at least four hydrogen-bonding and two hydrophobic interactions (Fig. 3a).

To identify protein sequences that not only buttress the ligand-coordinating residues from the RIF docking but also have low intra-protein energies to drive protein folding, we developed and applied a Monte Carlo-based sequence design protocol that iterates between (1) fixed-backbone design around the ligand-binding site to optimize the ligand-interaction energy and (2) flexible-backbone design for the rest of protein, optimizing the total complex energy (Fig. 3b). Forty-two designs with large computed folding-energy gaps and low-energy intra-protein and protein–ligand interactions were selected for experimental characterization, plus an additional 14 disulfide-bonded variants (Extended Data Fig. 5e). Ligand-docking simulations after extensive structure refinement revealed that owing to the approximate symmetry of the hydrogen-bonding pattern of DFHBI, many of the designed binding pockets could accommodate the ligand in two equally favourable orientations (Extended Data Fig. 5f).

Synthetic genes encoding the 56 designs were obtained and the proteins were expressed in *E. coli*. Thirty-eight of the proteins were well-expressed and soluble; SEC and far-ultraviolet CD spectroscopy showed that 20 were monomeric β-sheet proteins (Supplementary Table 3). Four of the oligomer-forming designs became monomeric upon incorporation of a disulfide bond between the N-terminal 3–10 helix and the barrel β-strands. The crystal structure of one of the monomeric designs (b10) was solved to 2.1 Å, and was found to be very close to the design model (0.57 Å backbone r.m.s.d., Fig. 3c). The upper barrel of the crystal structure maintains the designed pocket, which is filled with multiple water molecules (Fig. 3c, Extended Data Fig. 6b). Thus,

© 2018 Springer Nature Limited. All rights reserved.

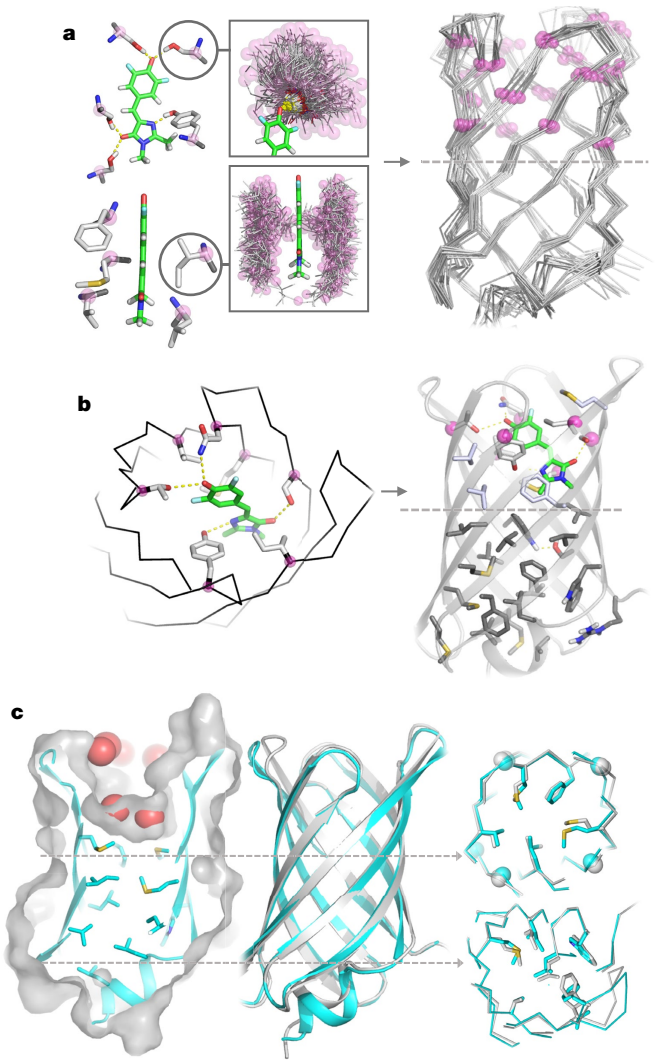


Fig. 3 | Computational design and structural validation of β-barrels with recessed cavities for ligand binding. a, Left, ensembles of side chains generated by the RIF docking method making hydrogen-bonding (upper left) and hydrophobic interactions (lower left) with DFHBI (green); pre-generated interacting rotamers are shown in grey with backbone C_α highlighted by magenta spheres. Right, ensemble of 200 β-barrel backbones with C_α atoms surrounding the binding cleft indicated by magenta spheres. **b**, Each ligand–scaffold pair (left) with multiple ligand-coordinating interactions from RIF docking is subjected to Rosetta energy-based sequence design calculations (right): positions around the ligand (light purple, above the dashed line) are optimized for ligand binding; the bottom of the barrel (dark grey, below the dashed line), for protein stability. **c**, Left, crystal structure (cyan cartoon with grey surface) of b10 with a recessed binding pocket filled with water molecules (red spheres). Middle, b10 design model backbone (silver) superimposed on the crystal structure (cyan). Right, comparison of crystal structure and design model for two different barrel cross-sections (indicated by dashed lines); glycine C_α atoms are indicated by spheres in the upper layer.

the design principles described above are sufficiently robust to enable the accurate design of potential small-molecule-binding pockets.

Two of the 20 monomeric designs—b11 and b32—were found to activate DFHBI fluorescence by 12- and 8-fold with binding dissociation constants (K_D) values of 12.8 and 49.8 μM, respectively (Extended Data Fig. 6f). Knockout of interacting residues in the designed binding pocket eliminated fluorescence (Extended Data Fig. 6g). The ligand-binding activity comes at a substantial stability cost as almost half of the barrel is carved out to form the binding site: although the nonfunctional BB1 design does not temperature-denature, both b11 and b32 undergo reversible thermal melting transitions (Extended Data Fig. 6e), even though b11 contains a stabilizing disulfide bond (the parent design that lacks the disulfide (b38) is not a monomer; Extended Data Fig. 6c, d). We sought to improve the binding interactions by redesigning β-turns around the ligand-binding site (Supplementary Table 6). b11L5F, a 110-residue protein with a five-residue fifth turn, activated DFHBI fluorescence by 18-fold with a K_D value of 7.5 μM (Extended Data Fig. 6f, h).

The sequence determinants of b11L5F fold and function were investigated by assaying the effect of each single amino acid substitution ($19 \times 110 = 2{,}090$ in total) on both protein stability[30] and DFHBI activation on the yeast cell surface. The function (fluorescence activation) and stability (proteolysis resistance) landscapes have similar overall features consistent with the design model, with residues buried in the designed β-barrel geometry being much more conserved than surface-exposed residues (Fig. 4a, Extended Data Fig. 7a, b). The function landscape suggests that the geometry of the designed cavity is critical to activating DFHBI fluorescence: the key sequence features that specify the geometry of the cavity—the glycine kinks and the tryptophan corner—are strictly conserved (Fig. 4a). Among all substitutions of the seven coordinating residues from RIF docking, only a single substitution (V103L) increased fluorescence (Fig. 4c, upper panel). Whereas the structure and function landscapes were very similar at the bottom of the barrel (Fig. 4b), there was a notable trade-off between stability and function at the top of the barrel around the designed binding site (Fig. 4c): many substitutions that stabilize the protein markedly reduce fluorescence activation (Fig. 4c, right). This bottom–top contrast indicates that success in de novo design of fold and function requires a substantial portion of the protein (in our case, the bottom of the barrel) to provide the driving force for folding as the functional site will probably be destabilizing.

Guided by the comprehensive protein stability and fluorescence activation maps, we combined substitutions at three positions that improved function without compromising stability (V103L, V95AG and V83ILM; Extended Data Fig. 8a, b), and obtained variants with ten-fold higher DFHBI fluorescence that form stable monomers without a disulfide bond (b11L5F.1; Extended Data Fig. 8c). The crystal structure of one of the improved variants (b11L5F_LGL; mutant 83L/95G/103L in Extended Data Fig. 8b) was solved to 2.2 Å and was very close to the design model with the majority of the buried side chains adopting the designed conformation (Extended Data Fig. 9a–d). However, the electron density in the binding site could not be resolved, consistent with the multiple DFHBI binding modes suggested by the docking calculations (Extended Data Fig. 9e–g; Extended Data Fig. 5f). A second round of computational design calculations was carried out to favour a specific binding mode by optimizing the protein–ligand interactions in the lowest-energy docked conformation, and to rearrange the hydrophobic packing interactions in the bottom of the barrel now freed from the disulfide bond. Five designs predicted by ligand-docking calculations to have a single ligand-binding conformation were experimentally tested and three showed increased fluorescence activity, the best of which increased the fluorescence by approximately 1.4-fold (b11L5F.2; Extended Data Fig. 8d–e). Screening of two combinatorial libraries (based on b11L5F.1 and b11L5F.2), incorporating additional beneficial substitutions identified in the b11L5F stability and function maps, yielded variants with another 1.5–2-fold increased fluorescence and improved protein stability (Extended Data Figs. 8f–h, 10a, b). We refer to these mini-fluorescence-activating proteins as mFAPs in the remainder of the text; mFAP0 and mFAP1 are variants of b11L5F.2, and mFAP2 of b11L5F.1. mFAP1 and mFAP2 activate 0.5 μM DFHBI fluorescence by 80- and 60-fold with K_D values of 0.56 μM and 0.18 μM, respectively (Fig. 5d).

The 1.8 Å and 2.3 Å crystal structures of mFAP0 and mFAP1 in complex with DFHBI were virtually identical to the design models with an overall backbone r.m.s.d. of 0.91 Å and 0.64 Å (Fig. 5a–c, Extended Data Fig. 9h, i). DFHBI is in the planar *Z* conformation with unambiguous electron density in both structures (Fig. 5a, Extended Data

© 2018 Springer Nature Limited. All rights reserved.

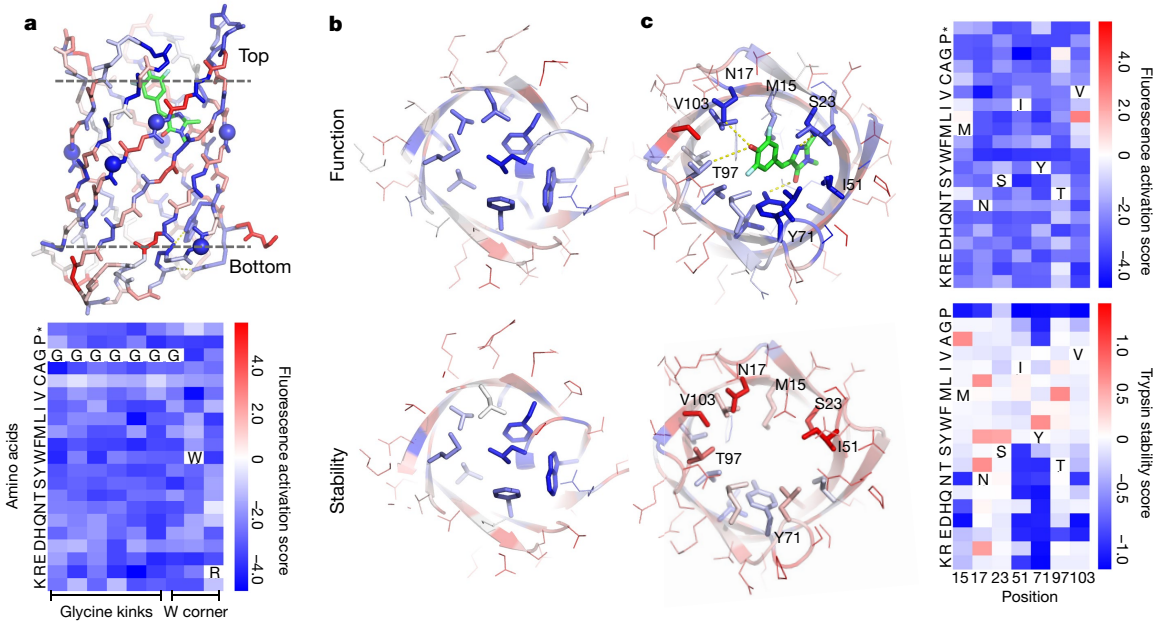


Fig. 4 | Sequence dependence of fold and function. Each position was mutated one at a time to each of the other 19 amino acids, and the resulting library was subjected to selection for fluorescence or stability to proteases. **a**, Top, b11L5F backbone model coloured according to relative fluorescence activation score at each position. Blue positions are strongly conserved during yeast selection; red positions are frequently substituted by other amino acids. Residues buried in the design model are much more conserved than solvent-exposed residues. Bottom, all the mutations to the glycine kinks (spheres in the top panel) and tryptophan (W) corner considerably reduced fluorescence; the shape of the designed structure is critical for the designed function. **b**, **c**, Bottom (**b**) and top (**c**) comparisons of b11L5F side chains coloured by relative fluorescence activation scores (top row) and stability scores (bottom row). In the bottom of the barrel, core residues were strongly conserved in both the function and stability selections (**b**); in the top barrel there is a clear function–stability trade-off with the key DFHBI-interacting residues being critical for function but far from optimal for stability (**c**, substitution patterns at these positions are shown on the right). Fluorescence activation and stability scores were derived from two biologically independent experiments with a greater than tenfold sequencing coverage. Standard deviation and confidence intervals are provided in Extended Data Fig. 7.

Fig. 9j). In addition to three designed hydrogen bonds, a water molecule was found to interact with the solvent-exposed phenol group in DFHBI (Fig. 5b). The DFHBI-binding modes in the crystal structures are nearly identical to the lowest-energy docked conformations used in the second round of design calculations, with all-atom r.m.s.d. values of 0.12 Å and 0.35 Å, respectively (Fig. 5c, Extended Data Fig. 9k). Three mutations shared by mFAP0 and mFAP1 in the bottom barrel (P62D, M65L and L85M or L85Y, Extended Data Fig. 10b) probably stabilize the protein by helical capping and subtle hydrophobic rearrangements (Extended Data Fig. 9l). The M27W mutation in mFAP1 introduced an additional hydrogen bond to DFHBI that probably produces the 5-nm red-shift in its fluorescence spectra (Fig. 5d, Extended Data Fig. 10c, e). mFAP2, which is based on b11L5F.1, has a six-residue insertion in the seventh β-turn that is predicted to form multiple intra-loop hydrogen bonds (Extended Data Fig. 10b, right).

In vivo fluorescence activation

To determine whether the designed DFHBI-binding fluorescence-activating proteins function in living cells, we imaged mFAP1 and mFAP2 in *E. coli*, yeast and mammalian cells by conventional wide field epifluorescence microscopy and confocal microscopy. Both mFAP1 and mFAP2 activated fluorescence less than 5 min after addition of 20 μM DFHBI. Cytosolic expression of mFAPs in *E. coli* and mammalian cells resulted in clear fluorescence throughout the cells (Fig. 5e, Extended Data Fig. 10f). Yeast cells with mFAPs targeted to the cell surface displayed fluorescence in a thin region outside of the plasma membrane (Fig. 5f, Extended Data Fig. 10g). Fusion of the mFAPs to a mitochondria-targeting signal peptide and to the endoplasmic reticulum-localized protein sec61β resulted in fluorescence tightly localized to these organelles in both fixed (Fig. 5g, h) and living cells (Supplementary Videos). The quantum yields of mFAP1 and mFAP2 in complex with DFHBI are 2.0% and 2.1%, respectively (Extended Data Fig. 4g, comparable with Y-FAST:HBR[31]). The brightness of de novo mFAPs in complex with DFHBI is about 35-fold lower than that of eGFP; there is still considerable room for improving their fluorescence activity.

Conclusion

It is instructive to compare the structures of our designed fluorescence-activating proteins with those of natural fluorescent proteins (Fig. 6). Both are β-barrels, and have similar chromophores, but our designs have less than half the residues and narrower barrels connected with short β-turns (Fig. 6a). In both cases, specific protein–chromophore interactions reduce energy dissipation from intramolecular motions[32], but the hydrogen bonding and hydrophobic packing around DFHBI is different from GFP and tailored to the smaller and simpler β-barrel (Fig. 6b). The precise structural control enabled by computational design, together with the greater exposure of the chromophore, may prove useful for fluorescence-based imaging and sensing applications.

The comparison in Fig. 6 highlights the two primary advances in this paper: the first successful de novo design of a β-barrel, and the first full de novo design of a small-molecule-binding protein. The first advance required the elucidation of general principles for designing β-barrels, notably the requirement for systematic symmetry-breaking to enable hydrogen bonding throughout the barrel structure. These principles, identified by pure geometric considerations, coupled with computer simulations after the failure of the initial parametric design approach, are borne out by both the crystal structures and the sequence fitness landscapes. The second advance goes considerably beyond the design of ligand-binding proteins and catalysts to date, which has relied on repurposing naturally occurring scaffolds. The three-step approach described in this paper—first, identifying the basic principles required for specifying a general fold class; second, using these principles to generate a family of backbones with pocket geometries

© 2018 Springer Nature Limited. All rights reserved.

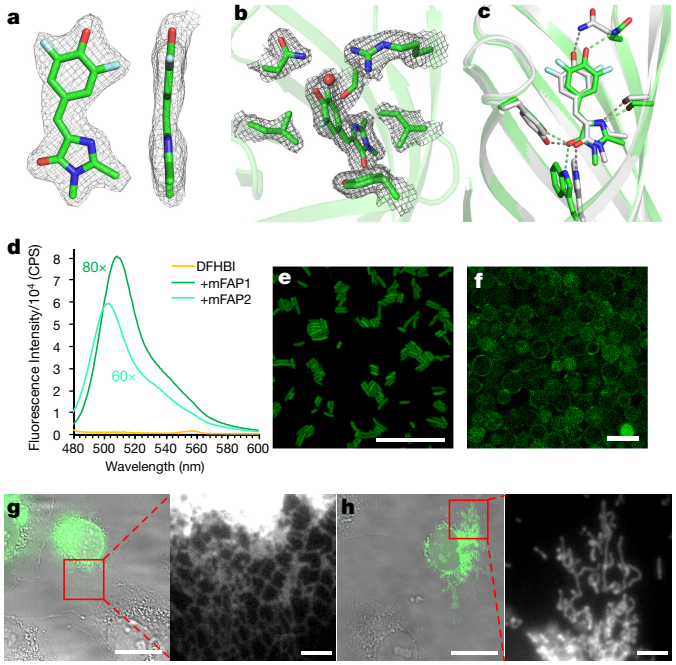


Fig. 5 | Structure and function of mFAPs. a, **b**, $2F_o - F_c$ omit electron density in the mFAP1–DFHBI complex crystal structure contoured at 1.0σ. **a**, DFHBI is clearly in the planar Z conformation rather than the non-fluorescent twisted conformations. **b**, The planar conformation is stabilized by closely interacting residues. **c**, Superposition of mFAP1 design model (silver) and the crystal structure (green). Hydrogen bonds coordinating DFHBI are indicated with dashed lines. **d**, Fluorescence emission spectra of 0.5 μM DFHBI with or without 5 μM mFAPs, excited at 467 nm. **e**, **f**, Confocal micrographs of *E. coli* cells expressing mFAP2 in the presence of DFHBI (**e**) and yeast cells displaying Aga2p–mFAP2 fusion proteins on the cell surface (**f**). Scale bars, 20 μm (**e**), 10 μm(**f**). **g**, **h**, Overlay of widefield epifluorescence (green) and brightfield (grey) images of fixed COS-7 cells expressing sec61β–mFAP1 (**g**) and mito–mFAP2 (**h**; mito, mitochondrial targeting sequence) with expanded views of the fluorescence in the boxed regions. Scale bars, 20 μm (**g**, **h**), 3 μm (expanded). Two biological replicates were performed with similar observation.

matched to the ligand or substrate of interest and third, designing complementary binding pockets buttressed by an underlying hydrophobic core—provides a general solution to the problem of de novo design of ligand-binding proteins. This generative approach enables the exploration of an effectively unlimited set of backbone structures with shapes customized to the ligand or substrate of interest and provides a test of our understanding of the determinants of folding and binding that goes well beyond descriptive analyses of existing protein structures.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0509-0.

Received: 19 March 2018; Accepted: 10 August 2018;
Published online 12 September 2018.

1. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
2. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
3. Tinberg, C. E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).
4. Bick, M. J. et al. Computational design of environmental sensors for the potent opioid fentanyl. *eLife* **6**, e28909 (2017).
5. Dou, J. et al. Sampling and energy evaluation challenges in ligand binding protein design. *Protein Sci.* **26**, 2426–2437 (2017).
6. Richardson, J. S. & Richardson, D. C. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).

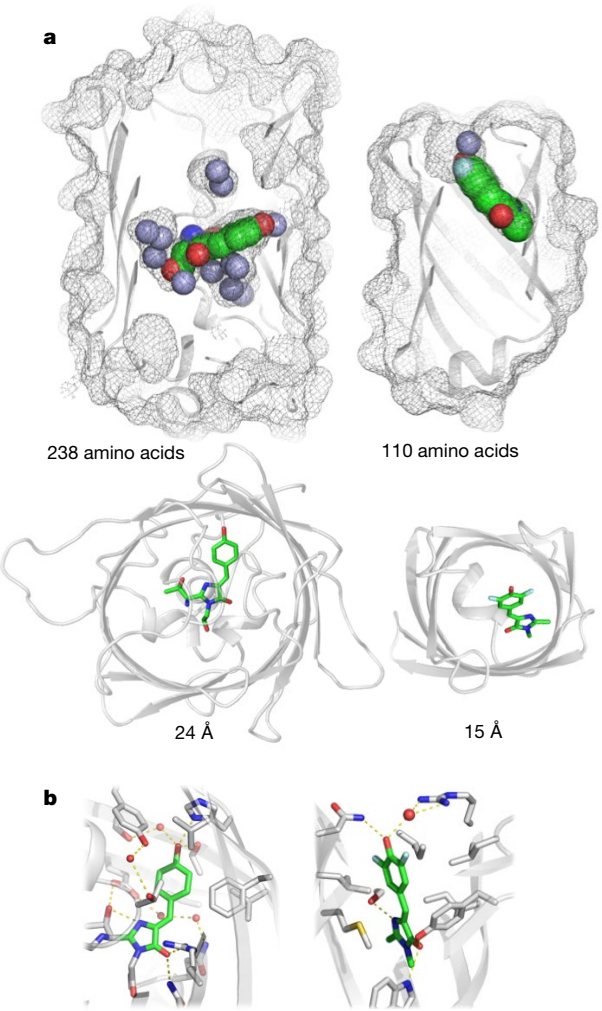


Fig. 6 | Comparison of structures of GFP and mFAP1. a, Surface mesh and ribbon representations of structures of GFP (left, PDB ID: 1EMA) and the computationally designed mFAP1 (right) with the chromophores (spheres) embedded in the protein. GFP, a product of natural evolution, has more than twice the number of residues, and a taller (top) and wider (bottom) barrel. Resolved water molecules in the crystal structures are shown as light purple spheres. **b**, Close-up of chromophore binding interactions in GFP (left) and mFAP1 (right).

7. Polizzi, N. F. et al. De novo design of a hyperstable non-natural protein–ligand complex with sub-Å accuracy. *Nat. Chem.* **9**, 1157–1164 (2017).
8. LaLonde, J. M., Bernlohr, D. A. & Banaszak, L. J. The up-and-down beta-barrel proteins. *FASEB J.* **8**, 1240–1247 (1994).
9. Flower, D. R. Structural relationship of streptavidin to the calycin protein superfamily. *FEBS Lett.* **333**, 99–102 (1993).
10. Richter, A., Eggenstein, E. & Skerra, A. Anticalins: exploiting a non-Ig scaffold with hypervariable loops for the engineering of binding proteins. *FEBS Lett.* **588**, 213–218 (2014).
11. Toda, M., Zhang, F. & Athukorallage, B. Elastic surface model for beta-barrels: geometric, computational, and statistical analysis. *Proteins* **86**, 35–42 (2018).
12. Novotný, J., Bruccoleri, R. E. & Newell, J. Twisted hyperboloid (Strophoid) as a model of β-barrels in proteins. *J. Mol. Biol.* **177**, 567–573 (1984).
13. Naveed, H., Xu, Y., Jackups, R., Jr & Liang, J. Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. *J. Am. Chem. Soc.* **134**, 1775–1781 (2012).
14. Lasters, I., Wodak, S. J., Alard, P. & van Cutsem, E. Structural principles of parallel beta-barrels in proteins. *Proc. Natl Acad. Sci. USA* **85**, 3338–3342 (1988).
15. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins. I. A theoretical analysis. *J. Mol. Biol.* **236**, 1369–1381 (1994).
16. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* **236**, 1382–1400 (1994).
17. Salemme, F. R. Conformational and geometrical properties of β-sheets in proteins. III. Isotropically stressed configurations. *J. Mol. Biol.* **146**, 143–156 (1981).

© 2018 Springer Nature Limited. All rights reserved.

18. Minor, D. L., Jr & Kim, P. S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367**, 660–663 (1994).
19. Fujiwara, K., Ebisawa, S., Watanabe, Y., Toda, H. & Ikeguchi, M. Local sequence of protein β-strands influences twist and bend angles. *Proteins* **82**, 1484–1493 (2014).
20. Lin, Y.-R. et al. Control over overall shape and size in de novo designed proteins. *Proc. Natl Acad. Sci. USA* **112**, E5478–E5485 (2015).
21. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
22. Richardson, J. S., Getzoff, E. D. & Richardson, D. C. The beta bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl Acad. Sci. USA* **75**, 2574–2578 (1978).
23. Chan, A. W., Hutchinson, E. G., Harris, D. & Thornton, J. M. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.* **2**, 1574–1590 (1993).
24. Hemmingsen, J. M., Gernert, K. M., Richardson, J. S. & Richardson, D. C. The tyrosine corner: a feature of most Greek key β-barrel proteins. *Protein Sci.* **3**, 1927–1937 (1994).
25. Greene, L. H., Hamada, D., Eyles, S. J. & Brew, K. Conserved signature proposed for folding in the lipocalin superfamily. *FEBS Lett.* **553**, 39–44 (2003).
26. Paige, J. S., Wu, K. Y. & Jaffrey, S. R. RNA mimics of green fluorescent protein. *Science* **333**, 642–646 (2011).
27. Warner, K. D. et al. Structural basis for activity of highly efficient RNA mimics of green fluorescent protein. *Nat. Struct. Mol. Biol.* **21**, 658–663 (2014).
28. Allison, B. et al. Computational design of protein–small molecule interfaces. *J. Struct. Biol.* **185**, 193–202 (2014).
29. Zanghellini, A. et al. New algorithms and an *in silico* benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).
30. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
31. Plamont, M.-A. et al. Small fluorescence-activating and absorption-shifting tag for tunable protein imaging in vivo. *Proc. Natl Acad. Sci. USA* **113**, 497–502 (2016).
32. Meech, S. R. Excited state reactions in fluorescent proteins. *Chem. Soc. Rev.* **38**, 2922–2934 (2009).
33. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

Acknowledgements We thank S. R. Jaffrey and T. A. Rapoport for providing experimental materials; A. Kang, S. A. Rettie, K. Lou, D. Sahtoe, D. La, G. J. Rocklin and C. Taylor for their help with experiments and data analysis; D. Alonso, L. Goldschmidt, P. Vecchiato, T. J. Brunette, D. Kim, V. K. Mulligan and T. Linsky for computer support, and the UW Hyak supercomputer and Rosetta@Home volunteers (https://boinc.bakerlab.org) for computing resources. We thank B. Huang, B. Basanta, R. Cacho, G. Daniel, Y. Kipnis, J. Klima and other members of the Baker laboratory for discussions. A.A.V. was supported by Fulbright Commission for Belgium and Luxembourg. E.M. was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2011-IOF 298976). B.L.S. is supported by NIH grant R01 GM115545. The Berkeley Center for Structural Biology is supported by the NIH, NIGMS and HHMI. The Advanced Light Source is a DOE User Facility under Contract No. DE-AC02-05CH11231. D.B. is supported by HHMI, WRF and Open Philanthropy.

Reviewer information *Nature* thanks R. Campbell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.A.V., J.D. and D.B. designed the study. W.S. developed RIF docking methods. B.M. developed the parametric design methods, designed and characterized the proteins. A.A.V. developed the β-barrel design methods with help from P.-S.H. L.C. purified proteins, performed SEC–MALS and analysed the results. L.A.D., M.J.B., B.S. and A.A.V. determined crystal structures. J.D. developed the ligand-binding design methodology and designed and optimized the mFAPs. H.P. performed post-design model refinement and docking calculations. G.W.F. and L.A.G. performed in vivo fluorescent imaging experiments. M.Y.L carried out photophysical characterization. E.M. and S.O. provided computational scripts. L.A.D. was supervised by B.L.S.; L.A.G. and M.Y.L were supervised by J.C.V. J.D., A.A.V. and D.B. wrote the manuscript with input from all authors.

Competing interests J.D., A.A.V. and D.B. are inventors on a U.S. provisional patent application submitted by the University of Washington that covers the described methods, sequences and applications.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0509-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0509-0.
Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Computational design of non-functional β-barrels. De novo design of non-functional β-barrels can be divided into two main steps: backbone construction and sequence design. For backbone construction, two different approaches were presented: parametric backbone generation and fragment-based backbone assembly. Example scripts and command lines for each method are available in Supplementary Data.

Parametric backbone generation and sequence design on the basis of hyperboloid models. β-strand arrangements were generated using the equation of a hyperboloid of revolution with an elliptic cross-section, sampling the elliptic radii around the ideal value of β-barrel radius with number of strands (n) and the shear number (S) (see Supplementary Methods). Eight β-strands were arranged as equally spaced straight lines running along the surface of the hyperboloid. A reference C_α atom was defined as the intersection between the first strand and the cross-section ellipse. The other C_α atoms were systematically populated along the eight strands from this reference residue. The peptide backbone was generated from the C_α coordinates using the BBQ software[34]. The arrangements of discrete β-strands were minimized with geometric constraints to favour backbone hydrogen bonds. One round of fixed-backbone sequence design calculation was carried out to pack the barrel cavity with hydrophobic residues. The resulting β-strand arrangements with the best hydrogen-bond connectivity and the tightest hydrophobic packing were selected to be connected by short (two to four residues) β-turns. Two iterations of the loop hashing protocol implemented in RosettaRemodel[35] were performed to close the strands and refine the turns. The sequence design of those β-turns was constrained to sequence profiles derived from natural proteins. Low-energy amino acid sequences were obtained for the connected backbones using a flexible-backbone design protocol (see Supplementary Data). Designs with high sequence propensity for forming β-strands, reasonable peptide-bond geometry and tight-packed hydrophobic cores are selected for experimental test (see Supplementary Table 2).

Backbone assembly from fragments guided by a 2D map. The presented 2D map (Fig. 1d) was designed with the longest strand length observed in soluble β-barrel structures to obtain a β-barrel tall enough to accommodate a hydrophobic core and a binding cavity. The length of each strand depends on its specific position and the shear number of the barrel (see Supplementary Methods). Glycine kinks and β-bulges were placed on the map as described in the main text. Specific β-turn types were used to connect the β-strands on the basis of their relative positions to β-bulges (see Supplementary Methods). On the basis of this 2D map, we generated a constraint file and a blueprint file to guide the assembly of the barrel using peptide fragments from Rosetta fragments library. In the constraint file, each backbone hydrogen bond was described as a set of distance and angle constraints (Extended Data Fig. 5b). A set of distance and torsion constraints specific to the tryptophan corner were added to the constraint file (Extended Data Fig. 3g–j, Supplementary Methods). In the blueprint file, a torsion angle bin was attributed to every residue in the peptide chain, according to the ABEGO nomenclature of Rosetta. After minimizing the assembled backbones using Rosetta centroid-scoring function with imposed constraints, our protocol output an ensemble of poly-valine β-barrel backbones with defined glycine kinks, β-bulges, β-turns and the backbone of the tryptophan corner. The main challenge of building scaffolds with this protocol is to properly balance structure diversity and reasonable backbone torsion angles with the strong geometric constraints imposed during minimization. For this work, we circumvented this problem by performing two additional rounds of sequence design calculation to regularize and prepare scaffolds for designing ligand-binding β-barrels (Extended Data Fig. 5b–d, Supplementary Methods).

Sequence design of nonfunctional β-barrels. Five hundred poly-valine backbones with good hydrogen bonds and torsion angles were selected as input for Rosetta sequence design. Low-energy sequences for the desired β-barrel fold were optimized over several rounds of flexible-backbone sequence design. We used a genetic algorithm to effectively search the sequence space: each parent backbone was used as input to produce ten designs through individual Monte Carlo searching trajectory. The best ~10% of the output designs were selected on the basis of the evaluation for total energy, backbone hydrogen bonds, backbone omega and Φ/Ψ torsion angles and hydrophobic-packing interactions. The selected models were used as inputs for the next round of design calculation. After 12 rounds of design and selection, no more improvements on the backbone quality metrics were observed (an indication of searching convergence). We then performed a backbone refinement by minimization in Cartesian space and a final round of design calculation (backbone flexibility was limited in torsion space for all the design calculation). The final top designs converged to the offspring of three initial backbones, sharing 36% to 99% sequence identity. For every parent backbone, one or two designs with the best hydrophobic packing interactions were selected for experimental characterization. The four designs (BB1–4) share 46% to 72% sequence identity.

Computational design of DFHBI-binding fluorescence-activating β-barrels. DFHBI is short for the chemical name ((Z)-4-(3,5-difluoro-4-hydroxybenzylidene)-1,2-dimethyl-1H-imidazol-5(4H)-one). De novo design of DFHBI-binding β-barrels consists of three steps: (1) generation of ensembles β-barrel scaffolds (see above), (2) ligand placement by RIF docking and (3) sequence design. Two hundred input scaffolds were generated in step 1 and used in the following steps. Example scripts and command lines are available in Supplementary Data.

RIF docking. The RIF docking method performs a simultaneous, high-resolution search of continuous rigid-body docking space as well as a discrete sequence-design space. The search is highly optimized for speed and in many cases, including the application presented here, is exhaustive for given scaffold–ligand pair and design criteria. RIF docking comprises two steps. In the first step, ensembles of interacting discrete side chains (referred to as 'rotamers') tailored to the target are generated. Polar rotamers are placed on the basis of hydrogen-bond geometry whereas apolar rotamers are generated via a docking process and filtered by an energy threshold. All the RIF rotamers are stored in ~0.5 Å sparse binning of the six-dimensional rigid body space of their backbones, allowing extremely rapid lookup of rotamers that align with a given scaffold position. To facilitate the next docking step, RIF rotamers are further binned at 1.0 Å, 2.0 Å, 4.0 Å, 8.0 Å and 16.0 Å resolution. In the second step, a set of β-barrel scaffolds is docked into the produced rotamer ensembles, using a hierarchical branch-and-bound search strategy (see Extended Data Fig. 5a). Starting with the coarsest 16.0 Å resolution, an enumerative search of scaffold positions is performed: the designable scaffold backbone positions are checked against the RIF to determine whether rotamers can be placed with favourable interacting scores. All acceptable scaffold positions (up to a configurable limit, typically ten million) are ranked and promoted to the next search stage. Each promoted scaffold is split into 2^6 child positions in the six-dimensional rigid-body space, providing a finer sampling. The search is iterated at 8.0 Å, 4.0 Å, 2.0 Å, 1.0 Å and 0.5 Å resolutions. A final Monte Carlo-based rotamer packing step is performed on the best 10% of rotamer placements to find compatible combinations.

Sequence design of DFHBI-binding β-barrels. A total number of 2,102 DFHBI–scaffold pairs from RIF docking were continued for Rosetta sequence design. Our design protocol iterated between a fixed-backbone binding-site design calculation and a flexible-backbone design for the rest of scaffold positions. Three variations of this design protocol were used during the sequence optimization. In the initial two rounds of design calculation, RIF rotamers (interacting residues placed during RIF docking) were fixed to maintain the desired ligand coordination. Repacking of RIF rotamers was allowed in the final round of design calculation, assuming that the binding sites have been optimized enough to retain these interactions. A Rosetta mover that biases aromatic residues for efficient hydrophobic packing was added after the first round of design. A similar selection approach and Cartesian minimization as described for non-functional sequence design were used to propagate sequence search and refine the design models. Evaluations on ligand-binding interface energy and shape complementarity were added to the selection criteria. The final set of designs was naturally separated into clusters on the basis of their original RIF docking solutions. For each cluster, a sequence profile was generated to guide an additional two rounds of profile-guided sequence design. Forty-two designs from 22 RIF docking solutions (20 input scaffolds) were selected for experimental characterization (see Supplementary Table 3).

Post-design model validation and ligand-docking simulation. To validate the protein and ligand conformations of the selected designs, we performed ligand-docking simulations on refined apo-protein models. Protein model refinement was carried out on the unbound model of the designs by running five independent 10-ns molecular dynamics simulations with structural averaging and geometric regularization[36]. Then ligand-docking simulation was performed on this refined unbound model using RosettaLigand[37] using Rosetta energy function[38], allowing rigid-body orientation and intra-molecular conformation of the ligand as well as surrounding protein residues (both on side chains and backbones) to be sampled. The ligand-binding energy landscapes were generated by repeating 2,000 independent docking simulations.

Design of disulfide bonds. The disulfide bonds were designed between the N-terminal 3–10 helix and a residue on one of the β-strands on the opposite side to the tryptophan corner. The first six residues of the designs model were rebuilt with RosettaRemodel[35] and checked for disulfide bond formation using geometric criteria. Once a disulfide bond was successfully placed, the N-terminal helix was redesigned.

Redesign of β-turns for b11. Three β-turns (loops 3, 5 and 7) surrounding the DFHBI-binding site of b11 were redesigned to make additional protein–ligand contacts. A set of 'pre-organized' loops with high content of intra-loop hydrogen bonds and low B-factors were collected from natural β-barrel structures, and used as search template to build individual loop fragment library. Those custom libraries were used as input for RosettaRemodel to build an ensemble of loop

© 2018 Springer Nature Limited. All rights reserved.

insertions in the b11 design model bound to DFHBI. Two rounds of flexible-backbone design calculation were carried out to optimize ligand interface energy and shape complementarity using sequence profiles to maintain the template backbone hydrogen bonds. Designed loop sequences were validated in silico by kinematic loop closure[39] (KIC). Five hundred loop conformations were generated by independent KIC sampling and scored by Rosetta energy function. Thirty-six designs with improved ligand-interface energy, shape complementarity and converging conformation sampling were selected for experimental characterization (see Supplementary Data, Supplementary Table 6).

Redesign of β-barrel core and DFHBI-binding site for b11L5F.1. After releasing the disulfide bond in b11L5F, with ligand modelled in the lowest-energy docked conformation for b11L5F (see Extended Data Fig. 5f, right), we performed another round of design calculation to further optimize the β-barrel core packing and ligand-binding interactions. The design protocol was very similar to the one used before with fixed ligand–hydrogen-bonding residues from RIF docking. Five designs with 9–15 mutations after manual inspection were selected for experimental characterization.

Protein expression and purification. Genes encoding the non-functional β-barrel designs (41 from parametric design and 4 from fragment-base design) were synthesized and cloned into the pET-29 vector (GenScript). Plasmids were then transformed into BL21*(DE3) *E. coli* strain (NEB). Protein expression was induced either by 1 mM isopropyl β-D-thiogalactopyranoside (IPTG) at 18 °C, or by overnight 37 °C growth in Studier autoinduction medium. Cells were lysed either by sonication (for 0.5–1-l cultures) or FastPrep (MPBio) (for 5–50-ml cultures). Soluble designs were purified by Ni-NTA affinity resin (Qiagen) and monomeric species were further separated by Akta Pure fast protein liquid chromatography (FPLC)(GE Healthcare) using a Superdex 75 increase 10/300 GL column (GE Healthcare). Fifty-six genes encoding DFHBI-binding designs were synthesized and cloned into pET-28b vector (Gen9). Protein expression and purification were carried out in the same way.

Circular dichroism (CD). Purified protein samples were prepared at 0.5 mg/ml in 20 mM Tris buffer (150 mM NaCl, pH 8.0) or PBS buffer (25 mM phosphate, 150 mM NaCl, pH 7.4). Wavelength scans from 195 nm to 260 nm were recorded at 25 °C, 75 °C, 95 °C and cooling back to 25 °C. Thermal denaturation was monitored at 220 nm or 226 nm from 25 °C to 95 °C. Near-ultraviolet wavelength scan from 240 nm to 320 nm and tryptophan fluorescence emission were recorded in the absence and presence of 7 M guanidinium chloride (GuHCl). Chemical denaturation in GuHCl was monitored by both tryptophan fluorescence and near-ultraviolet CD signal at 285 nm. The concentration of the GuHCl stock solution was measured with a refractometer (Spectronic Instruments). Far-ultraviolet CD experiments were performed on an AVIV model 420 CD spectrometer (Aviv Biomedical). Near-ultraviolet CD and tryptophan fluorescence experiments were performed on a Jasco J-1500 CD spectrometer (Jasco). Protein concentrations were determined by 280 nm absorbance with a NanoDrop spectrophotometer (Thermo Scientific). Melting temperatures were estimated by smoothing the sparse data with a Savitsky–Golay filter of order 3 and approximating the smoothed data with a cubic spline to compute derivatives. Reported T_m values are the inflection points of the melting curves.

Size-exclusion chromatography with multi-angle light scattering. Protein samples were prepared at 1–3mg/ml and applied to a Superdex 75 10/300 GL column (GE Healthcare) on a LC 1200 Series HPLC machine (Agilent Technologies) for size-based separation, and a miniDAWN TREOS detector (Wyatt Technologies) for light-scattering signals.

Fluorescence binding assay. Protein-activated DFHBI fluorescence signals were measured in 96-well plate format (Corning 3650) on a Synergy neo2 plate reader (BioTek) with $\lambda_{ex} = 450$ nm or 460 nm and $\lambda_{em} = 500$ nm or 510 nm. Binding reactions were performed at 200 μl total volume in PBS pH 7.4 buffer. Protein concentrations were determined by 280 nm absorbance as described above. DFHBI (Lucerna) were resuspended in DMSO as instructed to make 100 mM stock and diluted in PBS to 0.5–10 μM.

Library construction. Deep mutational scanning library for b11L5F were constructed by site-directed mutagenesis as described[40]. One hundred and eleven PCR reactions were carried out using DNA oligonucleotides directed to each position in two 96-well polypropylene plates (USA Scientific, 1402-9700), and products were pooled and purified by gel extraction kit (Qiagen) for yeast transformation. Combinatorial libraries for b11L5F.1 and b11L5F.2 were assembled using synthesized DNA oligonucleotides (Integrated DNA technologies) as described[41]. Selected positions were synthesized with 1–2% mixed bases to control mutation rate and library size. Full-length assembled genes were amplified and purified for yeast transformation as described[42].

Yeast surface display and fluorescence-activated cell sorting. Transformed yeast cells (strain EBY100)[42] were washed and re-suspended in PBSF (PBS plus 1 g/l of BSA). DFHBI in DMSO stock was diluted in PBSF for labelling yeast cells at various concentrations. PBSF-treated cells were incubated with DFHBI for 30 min to 1 h at room temperature on a benchtop rotator (Fisher Scientific). Library selections were conducted using the GFP fluorescence channel at 520 nm with 488-nm laser on a SH800 cell sorter (Sony). Proteolysis treatment and fluorescence labelling were performed in the same way as described[30]. Cell-sorting parameters and statistics for all selections are given in Supplementary Table 16.

Deep sequencing and data analysis. Pooled DNA samples for b11L5F deep mutational scanning library were transformed twice to obtain biological replicates. Two libraries were treated and sorted in a parallel fashion. Yeast cells of naive and selected libraries were lysed and plasmid DNA was extracted as described[43]. Illumina adaptor sequences and unique library barcodes were appended to each library by PCR amplification using population-specific primers (see Supplementary Table 8). DNA was sequenced in paired-end mode on a MiSeq Sequencer (Illumina) using a 300-cycle reagent kit (Catalogue number: MS-102-3003). Raw reads were first processed using the PEAR program[44] and initial counts analysed with scripts adapted from Enrich[45]. Stability scores were modelled using sequencing counts from proteolysis sorts as described[30]. Unfolded states were modelled without disulfide bonds (cysteine replaced by serine). Function scores were modelled using sequencing counts from DFHBI fluorescence sorts. A simple meta-analysis statistical model with a single random effect was applied to combine two replicates using the framework developed in Enrich2[46].

BB1 crystal structure. BB1 protein was concentrated to 20 mg/ml in an AMICON Ultra-15 centrifugation device (Millipore), and sequentially exchanged into 20 mM Tris pH 8.0 buffer. The initial screening for crystallization conditions was carried out in 96-well hanging drop using commercial kits (Hampton Research and Qiagen) and a mosquito (TTP LabTech). With additional optimization, BB1 protein crystallized in 0.1 M BIS-Tris pH 5.0 and 2 M ammonium sulfate at 25 °C by hanging drop vapour diffusion with 2:1 (protein:solution) ratio. Diffraction data for BB1 was collected over 200° with 1° oscillations, 5-s exposures, at the Advanced Light Source (Berkeley) beamline 5.0.1 on an ADSC Q315R area detector, at a crystal-to-detector distance of 180 mm. The data was processed in space group $P2_1$ to 1.63 Å using Xia2[47]. The BB1 design model was used as a search model for molecular replacement using the program Phaser[48], which produced a weak solution (TFZ 6.5). From this, a nearly complete model was built using the Autobuild module in Phenix[49]. This required the rebuild-in-place function of autobuild to be set to 'False'. Iterative rounds of model building in the graphics program Coot[50] and refinement using Phenix.refine[51] produced a model covering the complete BB1 sequence. Diffraction data and refinement statistics are given in Supplementary Table18.

b10, b11L5F_LGL crystal structure and mFAPs–DFHBI co-crystal structures. b10 was initially tested for crystallization via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech). Crystallization conditions were then optimized in larger 24-well hanging drops. b10 crystallized in 100 mM HEPES pH 7.5 and 2.1 M ammonium sulfate at a concentration of 38 mg/ml. The crystal was transferred to a solution containing 0.1 M HEPES pH 7.5 with 3.4 M ammonium sulfate and flash-frozen in liquid nitrogen. Data were collected with a home-source rotating anode on a Saturn 944+ CCD and processed in HKL2000[52].

b11L5F_LGL was concentrated to 19.6 mg/ml (1.58 mM), incubated at room temperature for 30 minutes with 1 mM TCEP then mixed with an excess of DFHBI (re-suspended in 100% DMSO). b11L5F_LGL complexed with DFHBI was screened via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech) and crystallized in 100 mM Bis-Tris pH6.5 and 45% (v/v) Polypropylene Glycol P 400. The crystal was flash-frozen in liquid nitrogen directly from the crystallization drop. Data were collected with a home-source rotating anode on a Saturn 944+ CCD and processed in HKL2000[52].

mFAP0 and mFAP1 were mixed with excess DFHBI (re-suspended in 100% DMSO), while keeping the final DMSO concentration at less than 1%. The mFAP0 and mFAP1 complexes were then concentrated to approximately 41 mg/ml and 64 mg/ml, respectively, and initially tested for crystallization via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech). Crystallization conditions were then optimized in larger 24-well hanging drops macroseeded with poor-quality crystals obtained in sitting drops. mFAP0 complexed with DFHBI crystallized in 200 mM sodium chloride, 100 mM HEPES pH 7.5 and 25% (w/v) polyethylene glycol 3350. The crystal was transferred to the mother liquor plus 2 mM DFHBI and 10% (w/v) polyethylene glycol 400 then flash-frozen in liquid nitrogen. Data were collected at the Berkeley Center for Structural Biology at the Advanced Light Source (Berkeley), on beamline 5.0.2 at a wavelength of 1.0 Å. and processed in HKL2000[52]. mFAP1 complexed with DFHBI crystallized in 100 mM MES pH6.5 and 12% (w/v) polyethylene glycol 20,000. The crystal was transferred to the mother liquor plus 2 M DFHBI and 15% glycerol then flash-frozen in liquid nitrogen. Data were collected with a home-source rotating anode on a Saturn 944+ CCD and processed in HKL2000[52].

Structures were solved by molecular replacement with Phaser[48] via phenix[49] using the Rosetta design model with appropriate residues cut back to C_α and DFHBI removed. The structure was then built and refined using Coot[50] and

© 2018 Springer Nature Limited. All rights reserved.

phenix[51], respectively, until finished. Diffraction data and refinement statistics are given in Supplementary Table18.

Statistics and reproducibility. In Fig. 1c, the models were coloured on the basis of the mean values of repulsion energy by position (Rosetta fa_rep) derived from a set of poly-valine backbones relaxed with constraints ($n = 189$ independently generated models); relaxed with constraints with a glycine in the middle of each C_β-strip ($n = 186$ independently generated models) and relaxed without constraints with glycines and β-bulges ($n = 194$ independently generated models). This experiment has been performed twice on different sets of backbones and produced similar results. In Fig. 2b–e, BB1 was purified and sized with SEC at least five times independently, yielding different ratios of monomeric-to-oligomeric species (20–75%). The fraction of monomer could be increased by heat-shocking the cells at 42 °C shortly before induction. Two biological replicates of the far- and near-ultraviolet CD and tryptophan fluorescence spectra acquisition of BB1 were done with similar results, and the chemical denaturation experiment performed once. Extended Data Fig. 4a–c, the analysis of BB1 with SEC–MALS was repeated twice on independently prepared protein samples and similar molecular masses were obtained. Additionally, the experiments were repeated on one sample stored at 4 °C at different time points ($t = 0$, $t = 7$ days and $t = 30$ days); all experiments had similar results and confirmed the stability of the monomeric species. BB2, 3 and 4 were purified once. The molecular mass (with SEC–MALS) and the far-ultraviolet CD spectra of the purified proteins were tested once. The sizing of purified BB1 mutants was performed once, with wild-type BB1 as an internal control.

Cell culture and transfection. COS-7 cells (ATCC CRL-1651, confirmed negative for mycoplasma) were grown in DMEM supplemented with 1× NEAA, 100 units/ml penicillin, 100 μg/ml streptomycin, and 10% FBS; and collected using 0.25% trypsin EDTA. Per transfection, approximately one million cells were transfected with 2 μg of plasmid using 18 μl of Lonza s.e. cell supplement, 82 μl of Lonza s.e. nucleofection solution and pulse code DS-120 on a Lonza 4D X Nucleofector system. After nucleofection cells were immediately seeded into ibidi μ-Slide eight-well glass bottom chambers at a density of ~30,000 cells/well and incubated overnight at 37 °C.

Cell fixation. Cells were fixed at 37 °C for 10 min in PFA/GA fixation solution containing 100 mM aqueous PIPES buffer pH 7.0, 1 mM EGTA, 1 mM $MgCl_2$, 3.2% paraformaldehyde, 0.1% glutaraldehyde; reduced for 10 min with freshly prepared 10 mM aqueous sodium borohydride; then rinsed with PBS for 5 min.

Microscopy. Conventional widefield epifluorescence imaging was performed on an inverted Nikon Ti-S microscope configured with a 60 × 1.2 NA water-immersion objective lens (Nikon), a light emitting diode source (LED4D120, Thorlabs), a multiband filter set (LF405/488/532/635-A-000, Semrock) and images were captured with a Zyla 5.5 sCMOS camera (Andor). The samples were illuminated 470 nm light at an intensity of ~2 W/cm^2 and with 200 ms exposures. For live-cell experiments, samples were incubated at 37 °C with Gibco CO_2 Independent Medium containing 50 μM DFHBI for 10 min before imaging. Time-lapse movies were acquired over a period of 5 min with a 200-ms exposure every 5 s. For fixed cell imaging, samples were incubated at room temperature (~22 °C) in PBS containing 50 μM DFHBI for 10 min before imaging.

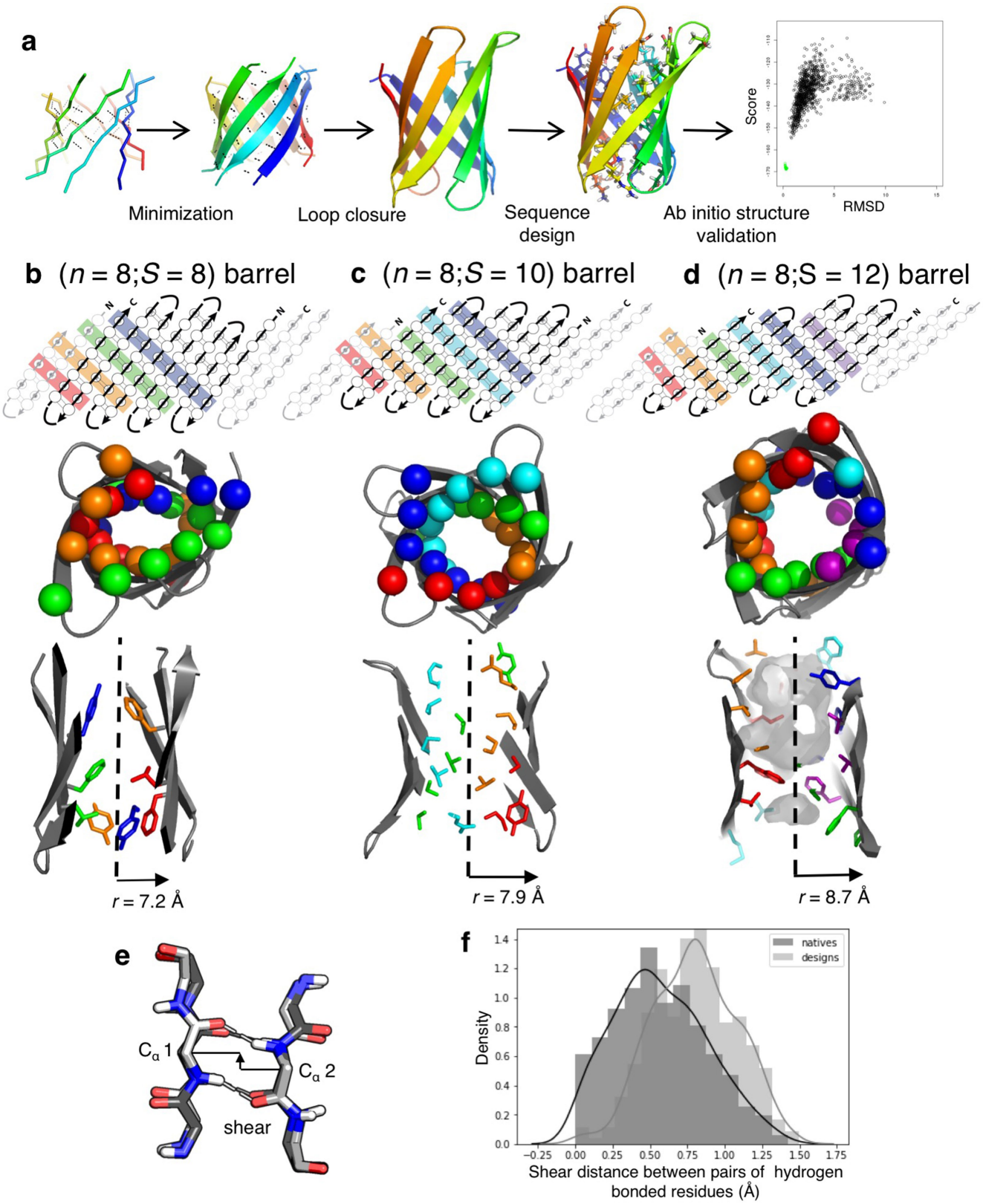
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The Rosetta macromolecular modelling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users. Commercial licenses for the suite are available via the University of Washington Technology Transfer Office. Design protocols and analysis scripts used in this paper are available in the Supplementary Information and at https://doi.org/10.5281/zenodo.1216229. The source code for RIF docking implementation is freely available at https://github.com/rifdock/rifdock.

Data availability. The atomic coordinates and experimental data of BB1, b10, b11L5F_LGL, mFAP0–DFHBI, and mFAP1–DFHBI crystal structures have been deposited in the RCSB Protein Database with the accession numbers of 6D0T, 6CZJ, 6CZG, 6CZH and 6CZI respectively. All the design models, Illumina sequencing data, sequencing analysis and source data (Figs. 2, 4, Extended Data Figs. 6e, 7, 8a, h) are available at https://dx.doi.org/10.5281/zenodo.1216229.

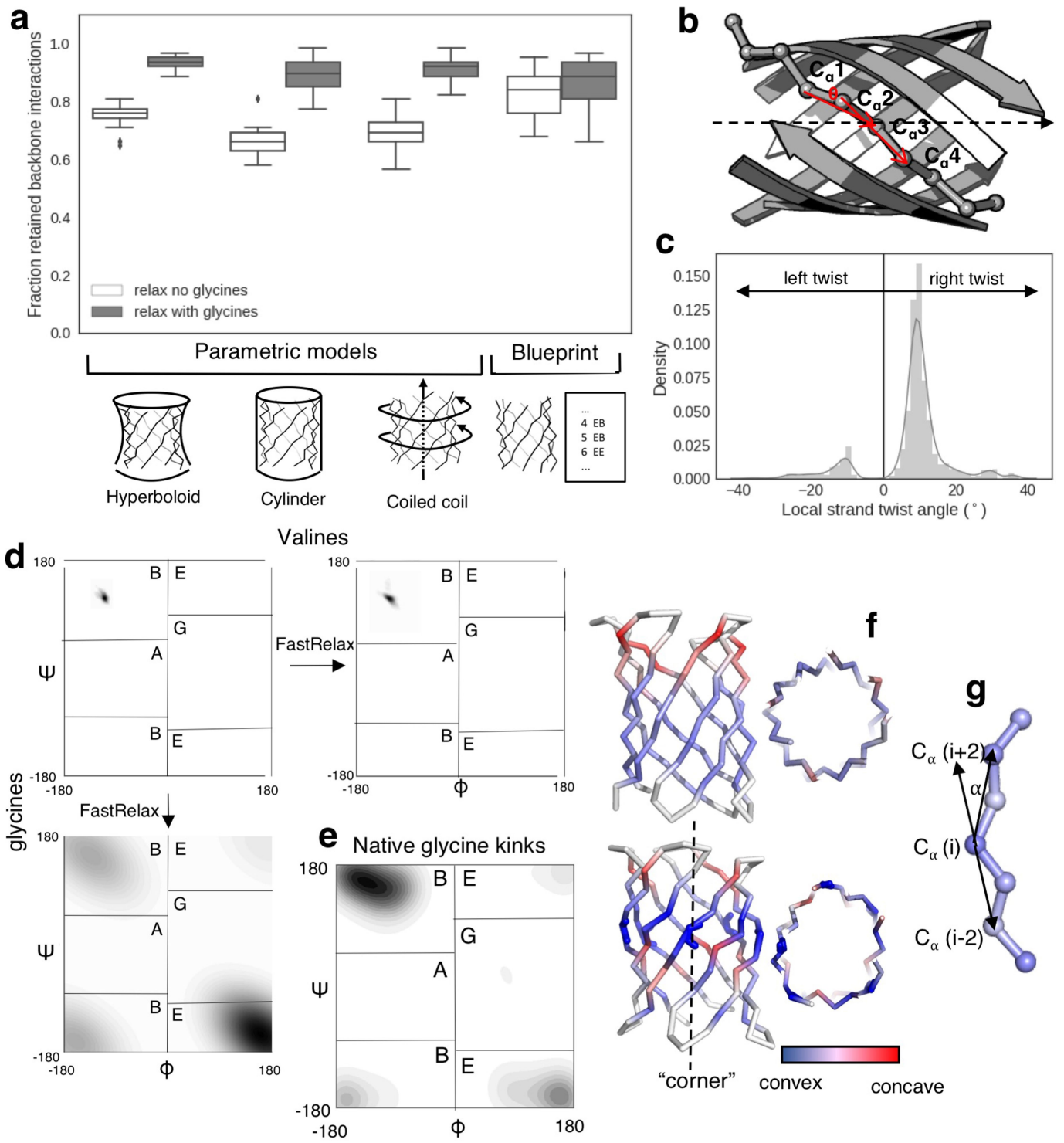
34. Gront, D., Kmiecik, S. & Kolinski, A. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.* **28**, 1593–1597 (2007).
35. Huang, P.-S. et al. RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6**, e24109 (2011).
36. Park, H., DiMaio, F. & Baker, D. The origin of consistent protein structure refinement from structural averaging. *Structure* **23**, 1123–1128 (2015).
37. Davis, I. W. & Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.* **385**, 381–392 (2009).
38. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
39. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6**, 551–552 (2009).
40. Procko, E. et al. Computational design of a protein-based enzyme inhibitor. *J. Mol. Biol.* **425**, 3563–3575 (2013).
41. Thyme, S. B. et al. Reprogramming homing endonuclease specificity through computational design and directed evolution. *Nucleic Acids Res.* **42**, 2564–2576 (2014).
42. Chao, G. et al. Isolating and engineering human antibodies using yeast surface display. *Nat. Protocols* **1**, 755–768 (2006).
43. Whitehead, T. A. et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).
44. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
45. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430–3431 (2011).
46. Rubin, A. F. et al. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).
47. Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).
48. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
49. Adams, P. D. et al. in *International Tables for Crystallography, Volume F*, 2nd edition (eds Arnold, E. et al.) Ch. 18.11, 539–547 (Wiley, Hoboken, 2012).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
51. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
52. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* Vol. 276 (Ed. Carter, W. C.) Ch. 3, 307–326 (Academic, Cambridge, 1997).
53. Merkel, J. S. & Regan, L. Aromatic rescue of glycine in β sheets. *Fold. Des.* **3**, 449–456 (1998).
54. Shaner, N. C., Steinbach, P. A. & Tsien, R. Y. A guide to choosing fluorescent proteins. *Nat. Methods* **2**, 905–909 (2005).
55. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).
56. Hauser, C. A. E. et al. Natural tri- to hexapeptides self-assemble in water to amyloid β-type fiber aggregates by unexpected α-helical intermediate structures. *Proc. Natl Acad. Sci. USA* **108**, 1361–1366 (2011).

© 2018 Springer Nature Limited. All rights reserved.



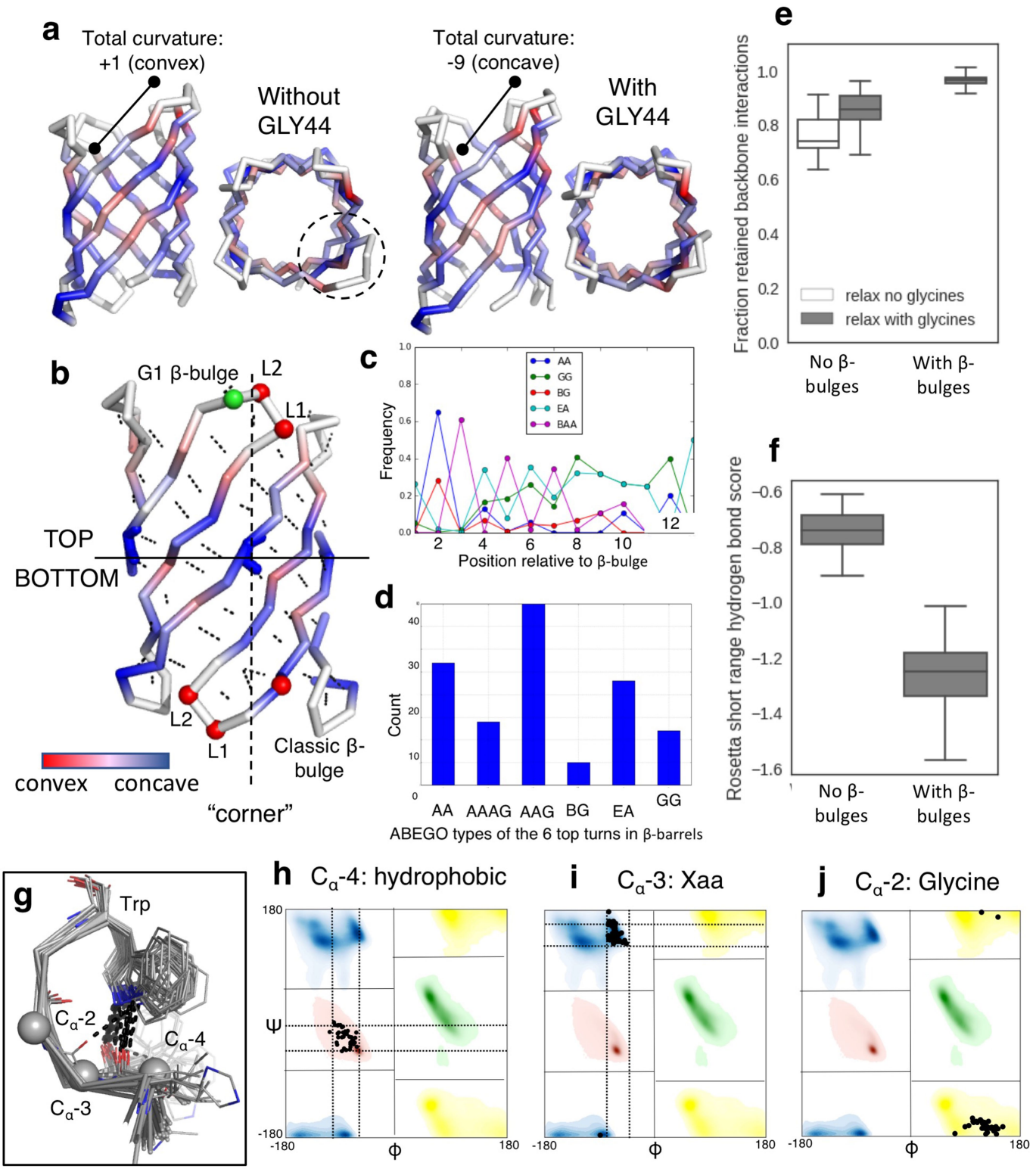
Extended Data Fig. 1 | Parametric design: workflow and shortcomings. **a**, Schematic of the parametric approach to generate β-barrel designs. **b**–**d**, Comparison between β-barrels of type ($n = 8$, $S = 8$) (**b**), type ($n = 8$, $S = 10$) (**c**) and type ($n = 8$, $S = 12$) (**d**); showing an example of 2D map with residue connectivity (top), the arrangement of the C_β atoms in the C_β-strips (middle) and the packing pattern of the core side chains (bottom). Although the three β-barrels have the same number of strands n, the difference in shear number S translates into different overall strand staggering, barrel radii (r) and number of C_β-strips. The number of core C_β-strips (top, middle) results in different arrangements of side chains in the core of the barrel. **e**, **f**, The parametric designs exhibited distorted hydrogen bonds, reflected by the shear distance (defined in **e**) between C_α atoms of paired antiparallel β-strand residues. The shear distance in the designs deviates from the distribution observed in native β-sheet proteins (**f**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | Glycine kinks release strain in β-barrel backbones. a, Fraction of retained hydrogen-bond interactions after relaxation with Rosetta ('relax') of uniform poly-valine backbones (white) and poly-valine backbones with a glycine in the middle of each C_β-strip (grey). We compare disconnected strand arrangements generated with the parametric hyperboloid model ($n = 225$ independently generated models), the cylindric model ($n = 36$ independently generated models), the coiled-coil model ($n = 150$ independently generated models) and assembled on the basis of a 2D map ($n = 144$ independently generated models). For all box plots: centre line, median; box limits, upper and lower quartiles; whiskers, minimum and maximum values; points, outliers. **b**, **c**, In poly-valine backbones ($n = 189$ independently generated models) relaxed with constraints to maintain hydrogen bonds between strands, several residues have unfavourable left-handed twist (**c**). The local strand twist is calculated on a sliding window of four residues along β-strands, as the angle between the vectors $C_\alpha1$–$C_\alpha3$ and $C_\alpha2$–$C_\alpha4$. The handedness of the twist is defined as the triple scalar product between these two vectors and the central axis of the barrel. Positive and negative values denote right-handed and left-handed twist, respectively (**b**). **d**, After relaxation ('FastRelax'), the valine positions in the middle of each C_β-strip remained in the β-sheet-specific ABEGO space (right); or were shifted towards the positive Φ space (E ABEGO) if mutated to glycines (bottom). **e**, A similar torsion angle distribution was observed for glycines in the β-strands of native β-barrels ($n = 35$ high-resolution crystal structures). **f**, In comparison with regular β-strands (top), the presence of glycine kinks (bottom) increases the local bending of the strands and creates corners in an otherwise circular barrel cross-section. **g**, The bending angle α is calculated on a sliding window of three residues.

© 2018 Springer Nature Limited. All rights reserved.

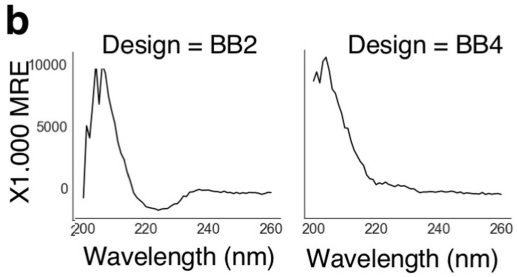


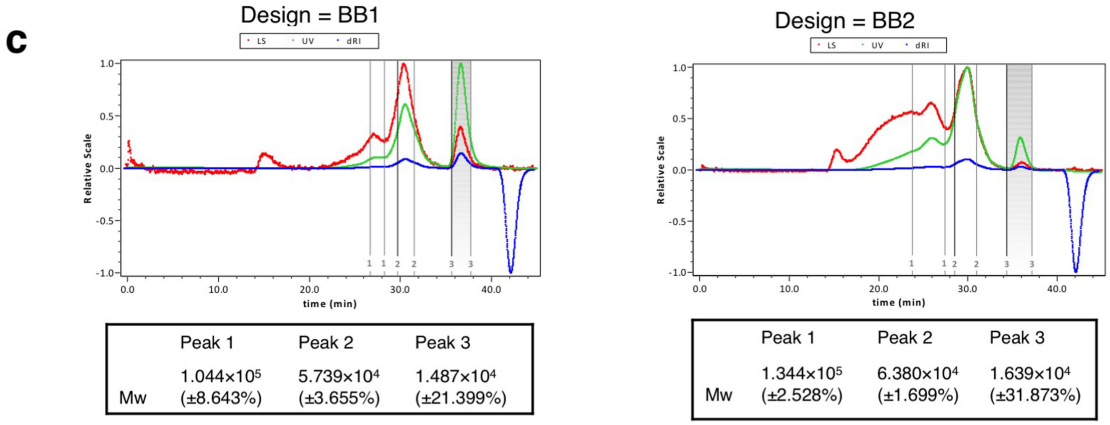
Extended Data Fig. 3 | Placement of β-bulges, β-turns and the tryptophan corner. a, Change of curvature (from convex to concave) and protrusion (dashed circle) of the longest hairpin associated with the placement of a glycine kink at position 44. **b**, Relationship between the 'corners' in the β-sheet (dashed line) generated by the glycine kinks and the type and position of the β-bulges and β-turns (Supplementary Methods). C_α are shown as spheres and coloured by ABEGO type. The bottom of the barrel was defined as the side of the N and C termini. **c**, The type I β-turn ('AA' ABEGO type) is frequently found at the second position relative to a β-bulge in native proteins and was selected to connect bottom hairpins. **d**, This choice is further supported by the enrichment of type I (AA) turns over the canonical type I′ turn (GG) in native β-barrels ($n = 35$ high-resolution crystal structures). **e**, **f**, Poly-valine backbones built with β-bulges and the corresponding β-turns ($n = 194$ independently generated models) retain more hydrogen bonds after relaxation than backbones built without β-bulges and with canonical type I′ β-turns ($n = 186$ independently generated models) (**e**) and exhibit better-scored hydrogen bonds per β-strand residue flanking the β-turns (**f**). **g**, Superposition of tryptophan corner motifs ($n = 41$ high-resolution crystal structures) extracted from native β-barrels. **h–j**, Amino acid preference and torsional constraints derived from the set and used to model the tryptophan corner. Bounded constraints limits are shown as dashed lines.

© 2018 Springer Nature Limited. All rights reserved.

a

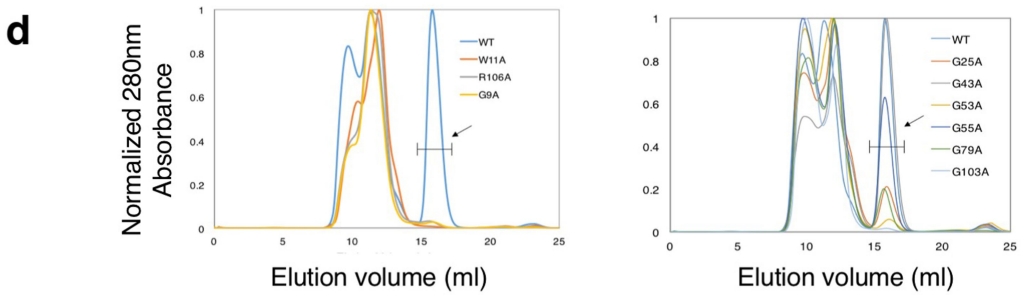
Design ID*	E-value†	*E.coli* Expression	Solubility	SEC	β CD spectrum
BB1	0.1100	yes	yes	**monomer**	yes
BB2	1.1000	yes	yes	tetramer	yes
BB3	1.9000	yes	no		
BB4	0.5600	yes	yes		no

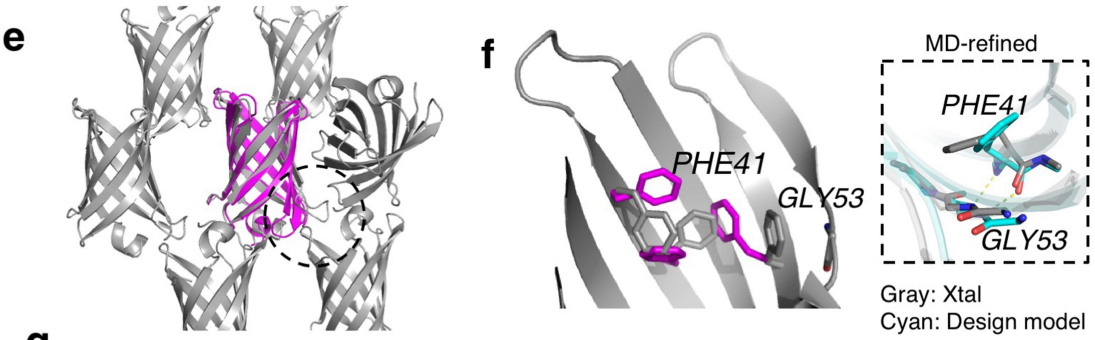




	Peak 1	Peak 2	Peak 3
Mw	1.044×10^5 (±8.643%)	5.739×10^4 (±3.655%)	1.487×10^4 (±21.399%)

	Peak 1	Peak 2	Peak 3
Mw	1.344×10^5 (±2.528%)	6.380×10^4 (±1.699%)	1.639×10^4 (±31.873%)





g

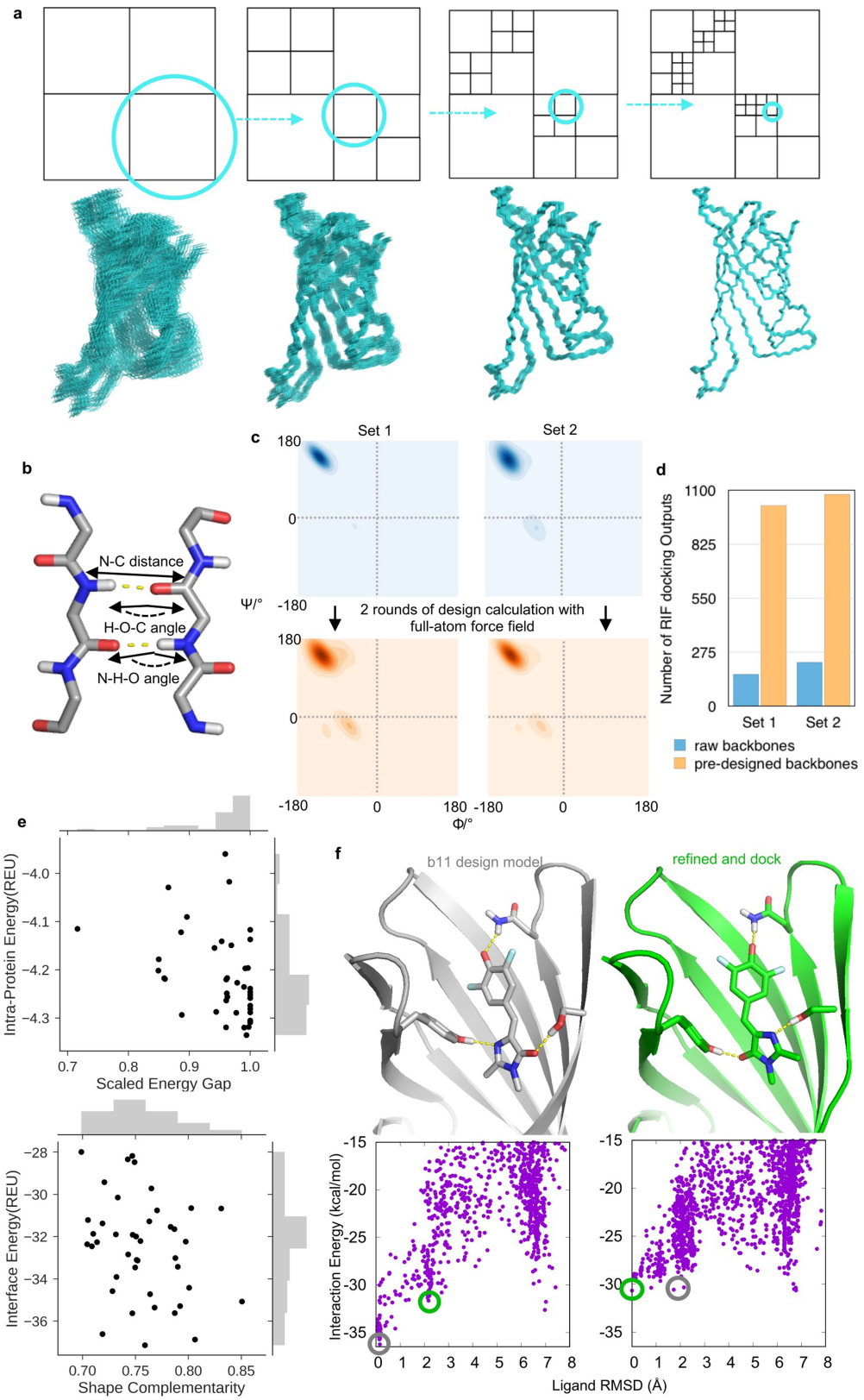
	λ_{abs}(nm)*	λ_{ex}(nm)*	λ_{em}(nm)*	Extinction Coefficient ($M^{-1}cm^{-1}$)	Brightness (extinction coefficient × QY)	Absolute Quantum Yield†	Relative Quantum Yield†	K_D(μM)
mFAP1-DFHBI	485	487	506	49,300	986 (34× dimmer than eGFP)	0.020	0.022	0.56 ± 0.07
mFAP2-DFHBI	476	480	500	43,200	907 (37× dimmer than eGFP)	0.021	0.021	0.18 ± 0.05
DFHBI	418	-	-	31,935	-	0.001	0.0007‡	
eGFP§		488	507	56,000	33,600		0.6	
Y-FAST:HBR‖	467		527	44,000	3,960		0.090	0.62 ± 0.05

Extended Data Fig. 4 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 4 | Biochemical and structural characterizations of designs BB1–4. a, Results of experimental characterization of the nonfunctional designs (BB1–4). Reproducibility is described in the Methods. †*E* value is calculated by BLAST, the non-redundant protein database. **b**, Far-ultraviolet CD spectra of designs BB2 and BB3 at 25 °C. **c**, SEC–MALS analysis showed a major monomer peak for BB1 and a major tetramer peak for BB2. **d**, Variants of BB1 with residues of the tryptophan corner and glycine kinks mutated to alanine were purified and sized. SEC traces are superimposed on the SEC trace of wild-type BB1 (WT). The mutations of all residues of the tryptophan corner eliminate the monomeric peak. Most of the glycine kink mutations negatively affect the monomeric species. The exceptions are Gly53 and Gly55, which are next to each other on the fourth strand. One glycine kink per strand might be sufficient to introduce enough negative twist to remove strain in the β-barrel. **e–f**, Deviations between BB1 design model and crystal structure. **e**, One of the three bottom turns of the crystal structure (grey) deviates from the design model (magenta) and forms additional crystal contacts (indicated by a dashed circle). **f**, Three phenylalanine side chains have different rotameric states. In the crystal structure, Phe41 interacts with Gly53 (which shows the most backbone deviation between the crystal structure and the design) to form an aromatic rescue motif[53]. It is likely that the discrepancies in the Phe rotamers reflect a scoring and sampling challenge to accurately capture such aromatic rescue; molecular dynamics simulation starting from the crystal structure (cyan) was also unable to recover the correct Phe41–Gly53 interaction. **g**, Biophysical properties (absorbance or fluorescence spectra, quantum yield and binding affinity) of mFAP1 and mFAP2 in complex with DFHBI. Mean values from three biological replicates were used for the nonlinear regression to determine the K_D. The error estimates are the standard deviation from the fitting calculation. *λ_{abs} is peak absorbance wavelength, λ_{ex} is peak excitation wavelength and λ_{em} is peak emission wavelength. †Absolute quantum yield is measured with an integrating sphere; relative quantum yield is measured using acridine yellow and fluorescein as the standards. ‡Previously reported value[26]. §Taken from previously published work[54]. ||Taken from previously published work[31].

© 2018 Springer Nature Limited. All rights reserved.

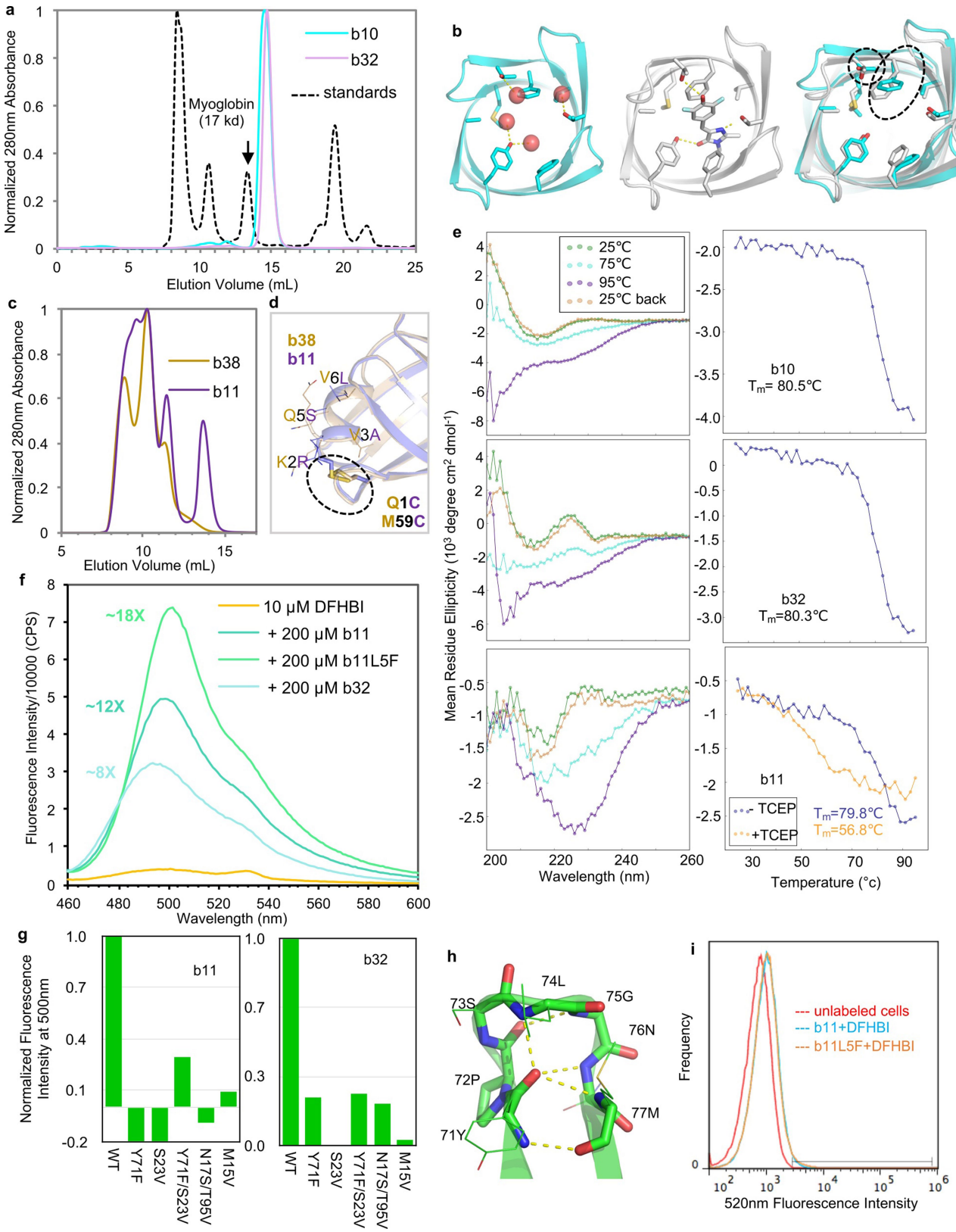


Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 5 | RIF docking grid-based search algorithm, β-barrel scaffold construction and post-design ligand-docking simulations. a, Illustration of grid-based hierarchical search strategy in RIF docking. After generating an ensemble of interactions for the target ligand (Fig. 3), each one of the selected scaffold is docked into the fixed RIF using the grid-based hierarchical searching algorithm. This search procedure starts from coarse sampling grids to fine sampling grids in 3D space. An example 2D grid scheme is shown in the upper row, from the lowest resolution (coarse sampling, left) to the highest resolution (fine sampling, right). At each searching stage, the backbone is assigned to different grids on the basis of its relative position and the resulting docking configurations are scored. The top-scored backbone positions (highlighted by cyan circles in the 2D scheme) are shown as 3D structures in the lower row for each searching resolution and are continued for the next grid search and scoring. The 3D structure example shown here was streptavidin structure (PDB ID: 1STP) with grid searching resolutions of 8.0 Å, 4.0 Å, 2.0 Å, and 1.0 Å. **b–d**, β-barrel scaffold construction for small-molecule binding. Three geometric constraints (**b**) were used to describe each backbone hydrogen bond and drive the backbone assembly during Rosetta low-resolution centroid modelling. Backbones generated with all three constraints had a very narrow Φ/Ψ distribution as a result of strong constraints (**c**, Ramachandran plot in upper left, set 1, density coloured in blue); by omitting N–H–O angle constraint, backbone torsion diversity slightly improved (**c**, upper right, set 2). These two raw backbone sets yielded few non-redundant RIF docking solutions (**d**, blue bars). After two rounds of sequence design calculation using Rosetta full-atom force field (Supplementary Methods), regularized backbones (peptide bonds with proper dihedral geometry) and broadened Φ/Ψ distribution (**c**, Ramachandran plot in the lower row, density coloured in orange) yielded more unique RIF docking solutions (**d**, orange bars). **e**, Computed metrics for 42 designs ordered and tested. Results from ab initio folding simulation were scaled to 0.0 to 1.0, in which 1.0 represents a funnel-shaped folding landscape[55]. **f**, Alternative ligand-binding conformations revealed by post-design ligand-docking simulations. The lowest-energy docking conformation using the design model (by simply taking out the ligand from the pocket) was similar to the designed DFHBI-binding mode (top left, grey; designed binding mode was circled in grey in the energy landscape in the lower row). Docking simulations using an apo-protein model refined by molecular dynamics simulations revealed an alternative equal-energy docking conformation (top right, green) that is indicated by a green circle in the docking energy landscapes (bottom). Both binding modes rely on three hydrogen-bonding residues from RIF docking (top).

© 2018 Springer Nature Limited. All rights reserved.

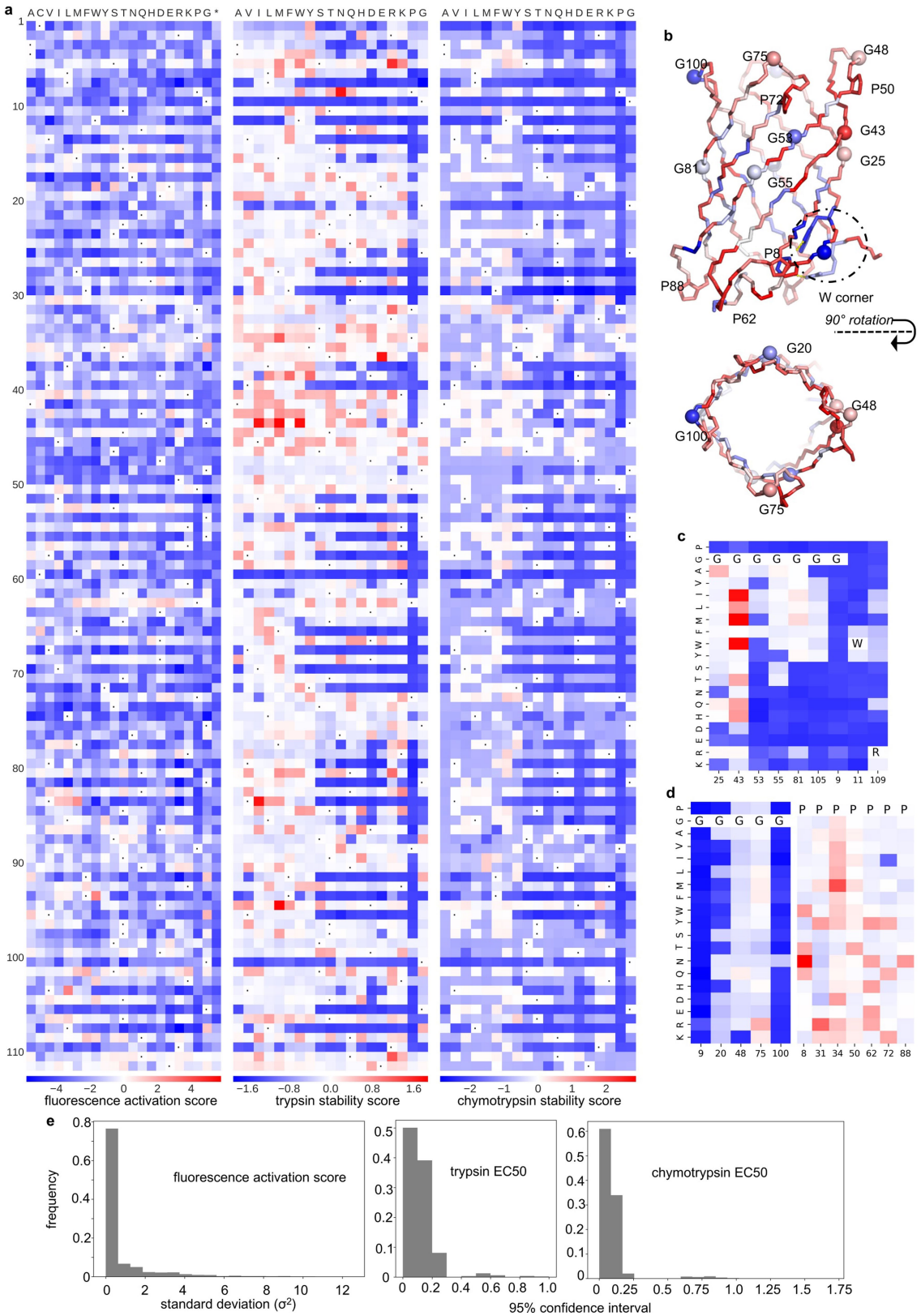


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | Biochemical and structural characterization of design b10, b32 and b11. a, Size-exclusion chromatogram of His6-tagged b10 and b32 after Ni-NTA affinity purification. The monodispersed peaks of absorbance at 280 nm of b10 and b32 (cyan and lavender, respectively) have an elution volume compatible with the monomeric β-barrel (14 kDa), on the basis of their relative position to the protein standard peaks (dashed line). Biological replicates were performed with similar observation: $n = 4$ for b10, $n = 5$ for b32. **b**, Comparison of the ligand-binding pocket in the b10 design model (middle, grey) with the crystal structure (left, cyan). The side chain disagreements are highlighted with a dashed black circle on the right panel. **c**, **d**, The designed disulfide bond as a stabilizing mechanism. SEC curves of His6-tagged b11 (purple line) and b38 (dark yellow line) were overlaid to show the appearance of a monomer peak for b11 (the same standard as in **a** was applied here). A disulfide bond connecting the N-terminal helix to a β-strand (Q1C and M59C, circled in **d**) and four mutations of neighbouring residues were introduced into design b38 (dark yellow) to make design b11 (purple). Biological replicates were performed with similar observation: $n = 3$ for b38, $n = 5$ for b11. **e**, Far-ultraviolet CD spectra of b10, b32 and b11. Left, spectra at different temperatures within one heating–cooling cycle. Right, thermal melting curves (the CD signal of b10 was monitored at 220 nm; b32 and b11 at 226 nm). b11 probably forms an amyloid-like beta structure at 95 °C (left, bottom row) with a negative peak around 226 nm[56] and refolds back after cooling to 25 °C. The thermal stability of b11 decreases when the disulfide was reduced with 1 mM tris(2-carboxyethyl) phosphine (TCEP) (right, bottom). Measurements were performed once for each design ($n = 1$). **f**, Fluorescence emission spectra of b32, b11 and b11L5F in complex with DFHBI. With 200 μM proteins, b32, b11 and b11L5F can activate 10 μM DFHBI fluorescence by 8-, 12- and 18-fold, respectively. Two biological replicates were performed with similar results. **g**, The residues designed to interact with DFHBI contribute to b11 and b32 activity. Single or double knockouts of hydrogen-bonding residues (Y71, S23, N17 and T95) and a hydrophobic-packing residue (M15) showed decreased fluorescence intensity at 500 nm in comparison with the wild-type b11 or b32 (WT). Mutants were purified once for activity measurement. **h**, **i**, Re-designed five-residue fifth turn in b11L5F. The original bulge-containing 'AAG' β-turn in b11 (Extended Data Fig. 3b) was redesigned into a five-residue turn. b11L5F was detected by yeast surface display and flow cytometry (**i** and Supplementary Data). Yeast cells displaying b11 and b11L5F showed an increased 520-nm fluorescence signal (excited by 488-nm laser, **i**). Three biological replicates were performed with similar observation.

© 2018 Springer Nature Limited. All rights reserved.

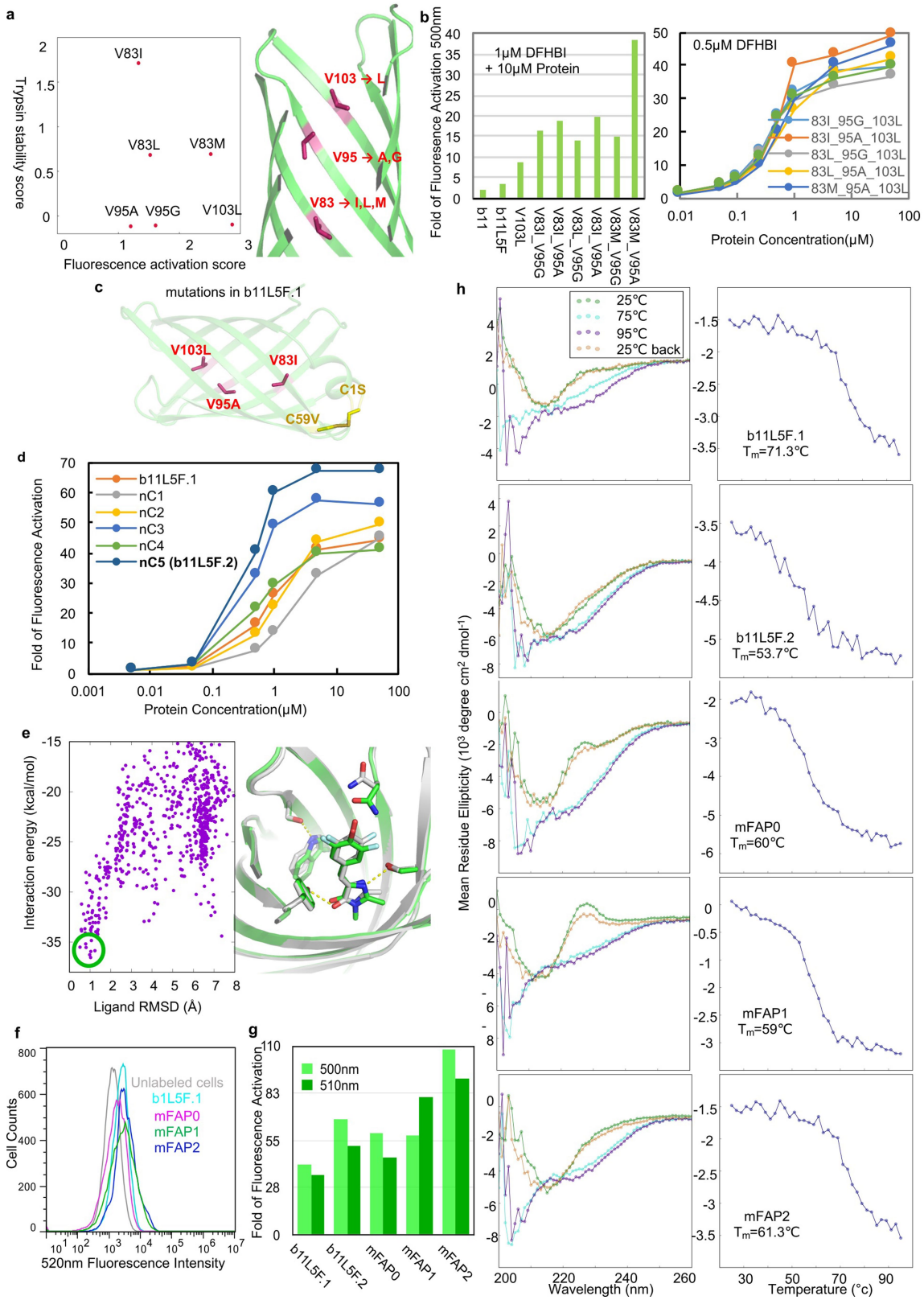


Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 7 | Deep mutational scanning maps for b11L5F. **a**, The complete function (left) and protease stability (middle and right) landscapes of b11L5F. Fluorescence activation scores, trypsin and chymotrypsin stability scores were calculated as described in Supplementary Methods and demonstrated in the Supplementary Data (b11L5F_DMS_analysis.ipy). Data are from two biological replicates with more than tenfold sequencing coverage. Red colour represents beneficial effect whereas mutations coloured in blue are detrimental (relative to the wild-type b11L5F). Wild-type residues at each position are indicated by black dots. **b**, b11L5F backbone model coloured one the basis of the average stability scores. Glycine backbone C_α are shown as spheres. **c**, **d**, Mutational scanning maps of glycine kinks (G25, G43, G53, G55, G81 and G105) and tryptophan corner positions (G9, W9 and R109) (**c**), and of glycines in the β-turns and prolines (**d**). **e**, Statistics of the fluorescence activation and stability scores. The standard deviation between the two replicates used for calculating fluorescence activation scores is smaller than two for most of the data points (left); 95% confidence interval calculated for the proteolysis/stability analysis is less than 0.25 for most the experimental protease half maximal effective concentration (EC_{50}) values (middle and right).

© 2018 Springer Nature Limited. All rights reserved.

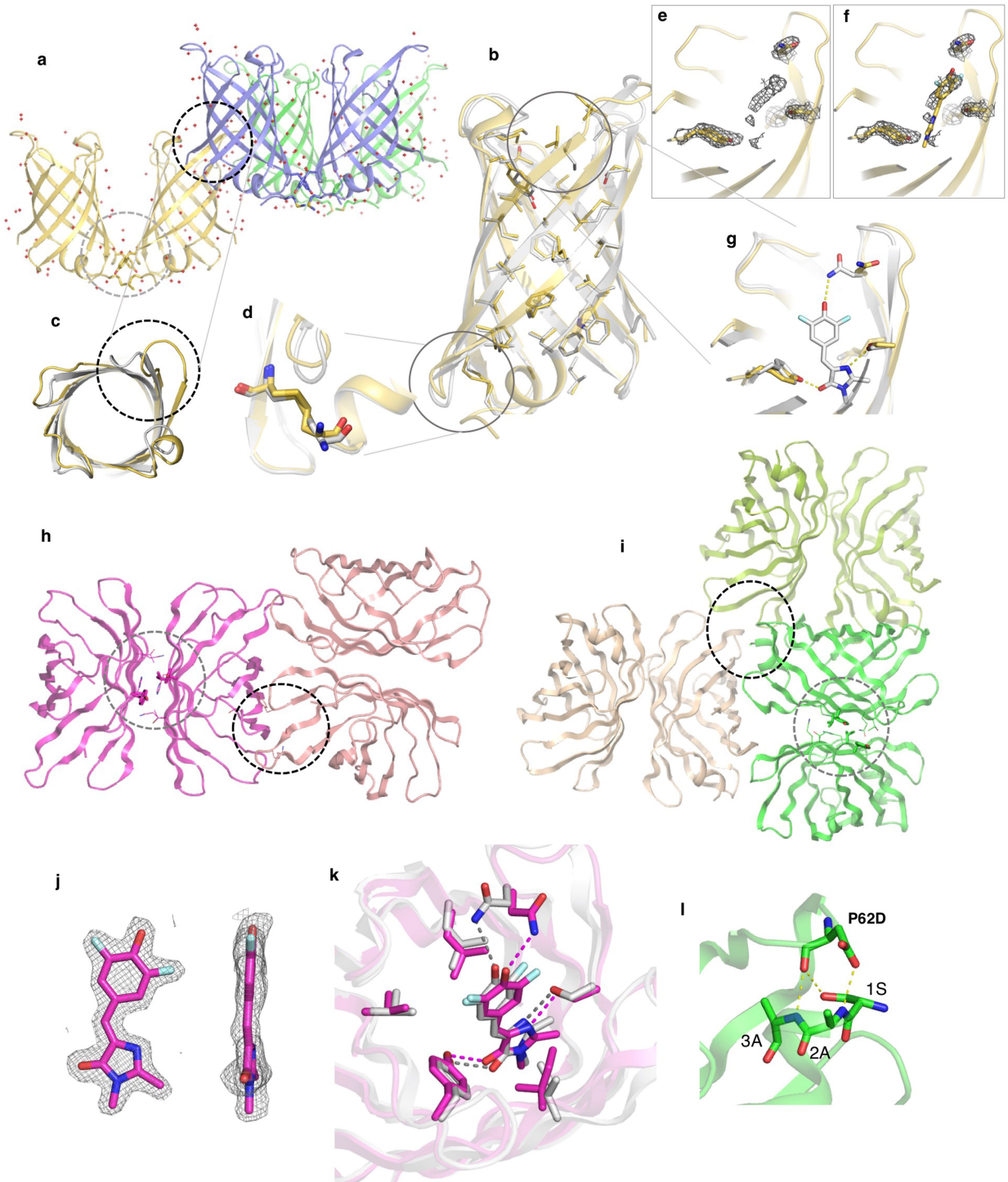


Extended Data Fig. 8 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

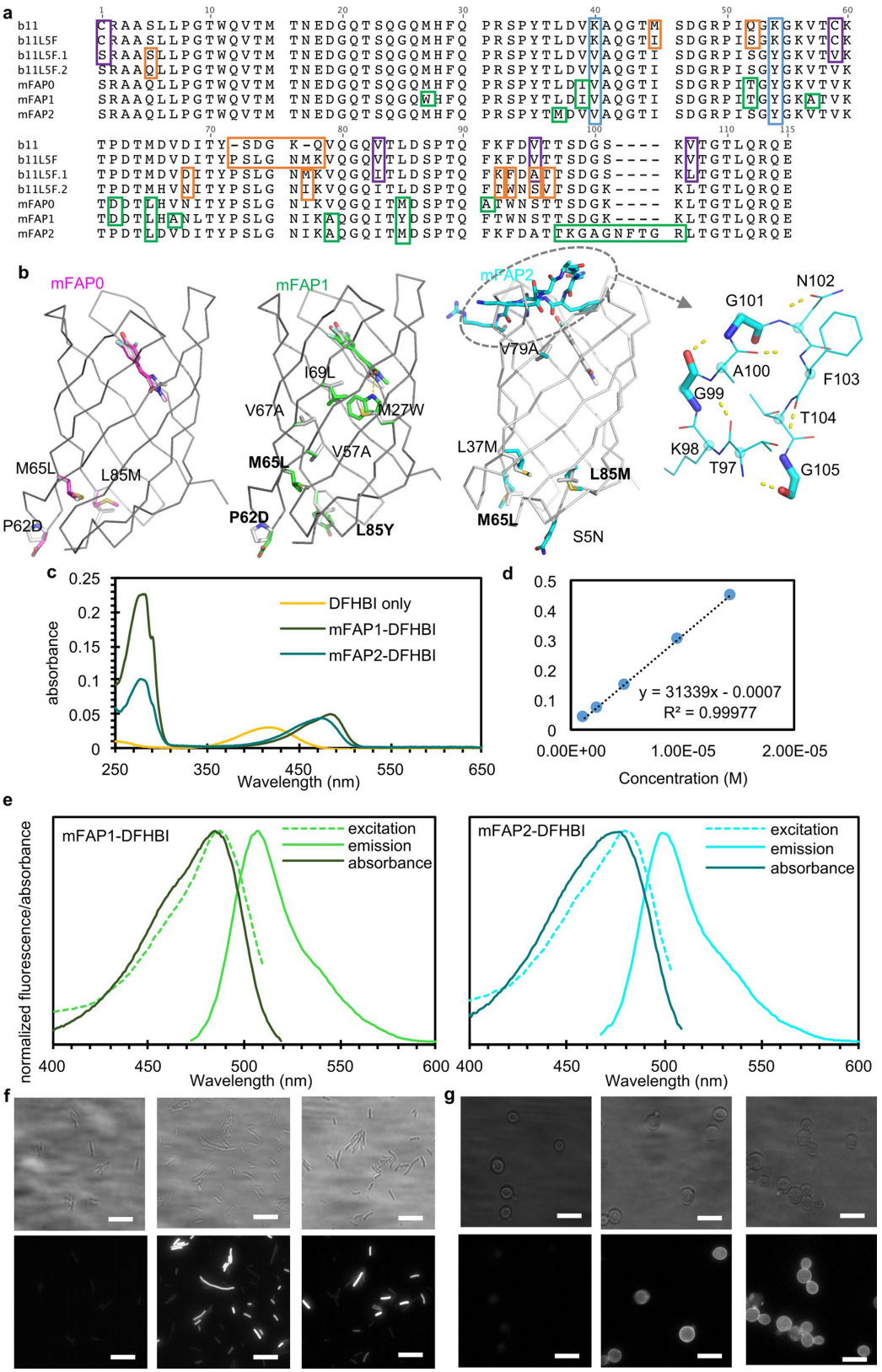
Extended Data Fig. 8 | Experimental and computational improvement on the basis of b11L5F. a–**c**, Incorporation of point mutations from deep mutational scanning. Beneficial mutations that improve fluorescence activity without compromising protein stability (positive scores relative to wild-type b11L5F; **a**, left, $n = 2$ biological replicates) were mapped onto b11L5F backbone model (**a**, right). **b**, Purified b11L5F variants incorporating those single, double or triple mutations showed consistently improved fluorescence activity. Binding titration curves were obtained for all six possible triple mutants (right, $n = 1$ biological measurement). **c**, b11L5F with V103L, V95A, V83I, C59V and C1S mutations was renamed as 'b11L5F.1'. **d**, Characterization of five designs from the second round of design calculation. Three of the five designs (nC1–5) that were based on b11L5F showed improved binding activities by titrating purified proteins into 0.5 μM DFHBI ($n = 1$ biological sample was used for the measurement). The best variant (nC5) was renamed 'b11L5F.2'. **e**, Ligand-docking simulations with the molecular dynamics-refined apo b11L5F.2. Energy landscape was plotted by comparing all the docking conformations to the design model (left). The lowest-energy docking conformations (highlighted in green circle) match the design model (right, design mode in silver and docking model in green). **f**, **g**, Characterization of three best variants (mFAP0–2) from combinatorial library selections. **f**, Yeast cells displaying mFAP proteins incubated with 5 μM DFHBI analysed by flow cytometry (excited by 488-nm laser, $n = 1$ biological sample was used for the measurement with proper controls). **g**, Purified proteins showed up to 100-fold fluorescence activation (5 μM protein + 0.5 μM DFHBI, excited at 450 nm and monitored at 500 nm and 510 nm in a plate reader, $n = 1$ biological measurement). **h**, Far-UV CD characterization of b11L5F.1, b11L5F.2, mFAP0, mFAP1 and mFAP2. Left, spectra at different temperatures within one heating–cooling cycle. Right, thermal melting curves (CD signals were monitored at 226 nm, spectra were recorded once ($n = 1$) with internal noise estimation).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Crystal structure of b11L5F_LGL, mFAP0 and mFAP1. a–g, b11L5F_LGL crystal structure. Protein samples of all six triple mutants in Extended Data Fig. 8b (right) were prepared for crystallization. b11L5F_LGL with V83L/V95G/V103L combination was successfully crystallized. Crystal contacts between protein copies in one asymmetric unit (yellow) were mediated by two tyrosines (stick representation, grey dashed circle); contacts between three asymmetric units (yellow, blue and green) were formed between β-turns (black dashed circle), which might have displaced one of the top β-turns (**c**). Overall backbone and side chain conformations in the design model matched the crystal structure with a backbone C_α r.m.s.d. of 1.02 Å (**b**, crystal in yellow and design model in silver), and the designed disulfide bond was present in the crystal structure (**d**). Ligand density in the crystal structure was ambiguous: $2F_o - F_c$ omit map showing the electron density after refinement without placing DFHBI (**e**), the best ligand placement to match the density (**f**), and designed ligand-binding interactions (silver) overlaid with the crystallized binding pocket (**g**). **h**, **i**, Crystal contacts in the DFHBI-bound structures of mFAP0 (**h**) and mFAP1 (**i**). Contacts between protein copies in one asymmetric unit were formed around 40V and 54Y (grey dashed circle) that were introduced for helping crystallization (Extended Data Fig. 10a). Contacts between asymmetric units were formed between β-turns (black dashed circle). **j**, $2F_o - Fc$ omit electron density of DFHBI in the mFAP0–DFHBI complex crystal structure. DFHBI density contoured at 1.0σ is clear and matches the planar conformation of the ligand (right). **k**, Superposition of mFAP0 design model (silver) and the crystal structure (magenta). Hydrogen bonds are indicated with dashed lines. **l**, Helical capping interactions mediated by P62D mutation in mFAP1 crystal structure.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | Mapping of mutations introduced into b11 to yield the final brighter variants, biophysical characterization of mFAP1 and 2, and epifluorescent images. a, Sequence alignment of b11-based DFHBI-binding fluorescence-activating proteins. Orange boxes indicate mutations or loop insertions introduced by computational design; purple boxes highlight mutations rationally introduced on the basis of the deep mutational scanning maps (Extended Data Figs. 7, 8); green boxes indicate mutations or loop insertions that were incorporated during combinatorial library selections; K40V and K54Y in light blue boxes were introduced to help crystal formation (Extended Data Fig. 9h, i). Despite having hydrophobic residues on the surface, mFAP2 remains soluble at 150 mg ml^{-1}. **b**, Mutations in the mFAPs mapped on the design models. Common mutations in all three mFAPs are highlighted in bold. **c**, Absorbance spectra for DFHBI, and the mFAP1–DFHBI and mFAP2–DFHBI complexes ($n = 4$ biological replicates with similar observations). **d**, Extinction coefficient determination for DFHBI at 418 nm. **e**, Normalized absorbance and fluorescence spectra of the mFAP1–DFHBI and mFAP2–DFHBI complex. Data are representative of two biological replicates with similar observations. **f, g**, Widefield epifluorescence (bottom) and brightfield (top) images of *E. coli* and yeast cells with 20 μM DFHBI. Untransformed *E. coli* Lemo21 cells (**f**, left, $n = 2$ biological replicates with similar observation) and yeast EBY100 cells displaying ZZ domain (**g**, left, $n = 2$ biological replicates with similar observation) were treated with the same amount of DFHBI and imaged in the same way (1000 mA 470-nm LED and 200-ms exposure time).

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0535-y

Cryo-EM structure of the active, G_s-protein complexed, human CGRP receptor

Yi-Lynn Liang[1,11], Maryam Khoshouei[2,10,11], Giuseppe Deganutti[3], Alisa Glukhova[1], Cassandra Koole[1], Thomas S. Peat[4], Mazdak Radjainia[1,5], Jürgen M. Plitzko[2], Wolfgang Baumeister[2], Laurence J. Miller[1,6], Deborah L. Hay[7,8], Arthur Christopoulos[1], Christopher A. Reynolds[3], Denise Wootten[1,9]* & Patrick M. Sexton[1,9]*

Calcitonin gene-related peptide (CGRP) is a widely expressed neuropeptide that has a major role in sensory neurotransmission. The CGRP receptor is a heterodimer of the calcitonin receptor-like receptor (CLR) class B G-protein-coupled receptor and a type 1 transmembrane domain protein, receptor activity-modifying protein 1 (RAMP1). Here we report the structure of the human CGRP receptor in complex with CGRP and the G_s-protein heterotrimer at 3.3 Å global resolution, determined by Volta phase-plate cryo-electron microscopy. The receptor activity-modifying protein transmembrane domain sits at the interface between transmembrane domains 3, 4 and 5 of CLR, and stabilizes CLR extracellular loop 2. RAMP1 makes only limited direct contact with CGRP, consistent with its function in allosteric modulation of CLR. Molecular dynamics simulations indicate that RAMP1 provides stability to the receptor complex, particularly in the positioning of the extracellular domain of CLR. This work provides insights into the control of G-protein-coupled receptor function.

CGRP is a physiologically important sensory neuropeptide with roles that include modulation of metabolism, inflammatory response and blood pressure, as well as auditory nerve development and function[1–4]. It is a potent vasodilator that is released during neurogenic inflammation and contributes to the pathology of migraine. A first-in-class drug targeting the CGRP receptor was recently approved for treatment of this condition, and many other therapeutic agents aimed at reducing CGRP activity are under development[5]. By contrast, CGRP is protective in models of inflammatory bowel disease and hypertension, and is a critical neuropeptide for development and modulation of auditory responses[1–4].

Receptor activity-modifying proteins (RAMPs) are essential accessory proteins for presentation of the class B CLR to the cell surface. They are integral components of the phenotypically ascribed CGRP and adrenomedullin receptors, through which CLR–RAMP1 mediates a selective response to CGRP, and CLR–RAMP2 or CLR–RAMP3 mediate selective responses to adrenomedullin[6]. RAMPs are also partners for the calcitonin receptor (CTR), but are not required for cell surface trafficking; they generate distinct amylin receptor phenotypes[1]. There is considerable crosstalk between calcitonin-family peptides and receptors; however, research has largely focused on how RAMPs influence cAMP signalling[1]. The three RAMPs each contain a structured, N-terminal extracellular domain (ECD) of about 100 amino acids, a single transmembrane domain and a short intracellular C terminus. There is evidence that RAMPs co-evolved with GPCRs[7]; supporting this, we and others have shown that they can partner with numerous GPCRs from all major subclasses, and are not exclusively partners for CLR and CTR[8–11].

Structures of heteromeric complexes of the isolated extracellular domains (ECDs) of RAMPs and CLR bound to C-terminal peptide fragments have been solved[12,13], and provide important but limited data on how RAMPs and CLR interact; however, they are unable to explain peptide selectivity. Therefore, structures of full-length, active CGRP and adrenomedullin receptor complexes are required.

Recent advances in cryo-electron microscopy (cryo-EM) have enabled elucidation of structures of full-length, class B GPCRs bound to peptide agonists, in complex with their canonical G_s-protein heterotrimers[14–16]. These studies revealed class-specific, conserved, global conformational changes linked to receptor activation, and unexpected divergence in the modes of peptide binding, even within the same receptor[14–18]. In the current work, we have used Volta phase-plate cryo-EM to determine the structure—at a global resolution of 3.3 Å—of the human CGRP receptor complex bound to its endogenous peptide agonist and canonical transducer. This structure provides insights into how RAMPs interact with GPCRs and modulate their activity.

Structure determination

We modified the CLR to replace the native signal peptide with that of haemagglutinin and affinity tags bracketed by 3C cleavage sites were introduced at the N and C terminus (Flag and His, respectively) (Extended Data Fig. 1). RAMP1 was modified with a haemagglutinin signal peptide followed by a Flag epitope (Extended Data Fig. 1). These modifications did not alter the receptor pharmacology (Extended Data Fig. 2a).

To form an active, G-protein-coupled complex, CLR and RAMP1 were co-expressed with $G\alpha_s$, His–$G\beta_1$ and $G\gamma_2$ in *Trichoplusia ni* insect cells, and stimulated with 10 μM CGRP. A stabilized $G\alpha_s$[15] was used together with the camelid antibody-derived nanobody Nb35[14–16], enabling formation of a complex with improved stability[19]. The complex was treated with 3C enzyme to remove tags from CLR, solubilized in lauryl maltose neopentyl glycol (LMNG) and cholesteryl hemisuccinate and then purified by sequential nickel-affinity and

[1]Drug Discovery Biology and Department of Pharmacology, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria, Australia. [2]Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, Martinsried, Germany. [3]School of Biological Sciences, University of Essex, Colchester, UK. [4]CSIRO Biomedical Manufacturing, Melbourne, Victoria, Australia. [5]Thermo Fisher Scientific, Eindhoven, The Netherlands. [6]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Scottsdale, AZ, USA. [7]School of Biological Sciences, University of Auckland, Auckland, New Zealand. [8]Maurice Wilkins Centre for Molecular Biodiscovery, University of Auckland, Auckland, New Zealand. [9]School of Pharmacy, Fudan University, Shanghai, China. [10]Present address: Novartis Institutes for Biomedical Research, Novartis Pharma, Basel, Switzerland. [11]These authors contributed equally: Yi-Lynn Liang, Maryam Khoshouei. *e-mail: denise.wootten@monash.edu; patrick.sexton@monash.edu

© 2018 Springer Nature Limited. All rights reserved.

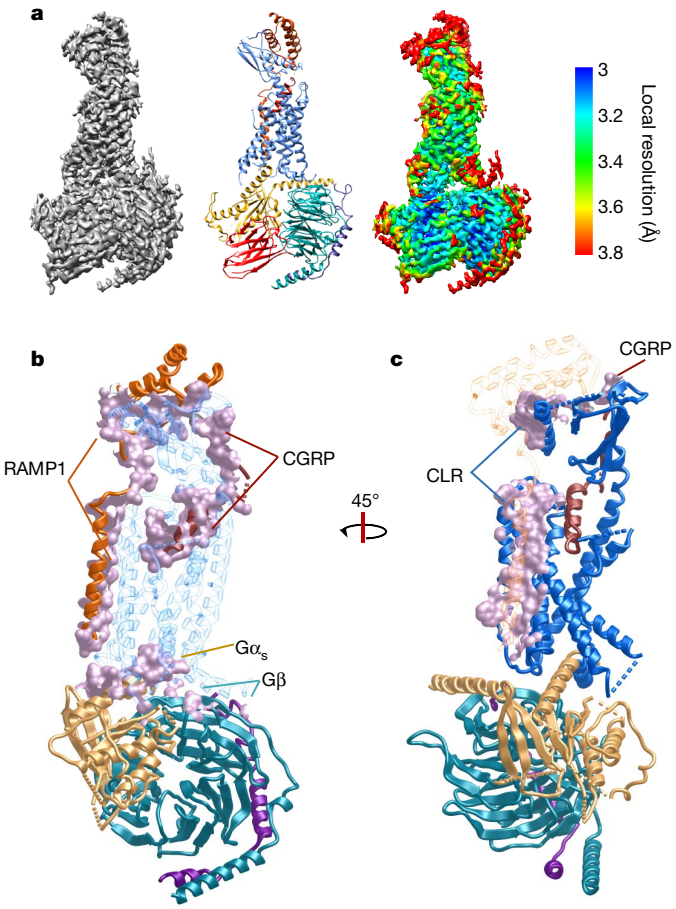


Fig. 1 | The cryo-EM structure of CGRP–CLR–RAMP1–G_s reveals molecular details of the RAMP-receptor interface. a, Left, 3.3 Å-resolution cryo-EM density map of the CGRP–CLR–RAMP1–G_s complex; the detergent micelle has been masked out for clarity. Middle, the structure in ribbon representation after refinement in the cryo-EM map. CGRP, dark red; CLR, blue; RAMP1, dark orange; $G\alpha_s$-Ras domain, gold; Gβ-subunit, cyan; Gγ-subunit, purple; Nb35, red. Right, the cryo-EM density map coloured by local resolution. **b**, **c**, CGRP receptor complex (ribbon representation coloured as in **a**), illustrating the extent of CLR interactions with other proteins in the complex (**b**), or the extent of RAMP1 interactions with other proteins in the complex (**c**), shown in mauve coloured surface representation. CGRP and RAMP1 form extensive contacts with CLR, with 61.5% and 23% of their surface being buried, respectively.

Flag-antibody columns, to ensure that only RAMP1-bound complexes were present, and then further purified by size-exclusion chromatography to yield a monodisperse complex that contained all the components (Extended Data Fig. 2b, c).

Vitrified complexes were imaged using a Titan Krios microscope equipped with a Volta phase plate[20,21]. Following imaging (Extended Data Fig. 3a) and initial 2D classification (Extended Data Fig. 3b), 3D classification yielded a final map at a resolution of 3.3 Å reconstructed from 407,000 particle projections (Fig. 1a, Extended Data Fig. 3c–e, Supplementary Table 1). The cryo-EM density map exhibited well-resolved side chains, allowing confident rotamer placements for most amino acids within the peptide, receptor and RAMP transmembrane domains, and the G protein (Extended Data Fig. 4). The RAMP and CLR ECDs had lower overall resolution, with discontinuous density for CLR ECD loop 1 and loop 5 (Fig. 1a, Extended Data Figs. 1, 5). Nonetheless, there was a strong correlation between the cryo-EM density of the ECD and those of the individual ECDs of either CLR or RAMP1 in a deposited X-ray structure (RCSB Protein Data Bank code (PDB): 4RWG[12]). These were rigid-body fitted into the ECD density, with side-chain adjustment where this was supported by density in the cryo-EM map. Whereas individual ECDs from the X-ray structures exhibited close approximation to the cryo-EM map, there were differences between the two structures in the relative positioning of the CLR and RAMP1 ECDs (Extended Data Fig. 5) that are likely to have arisen from anchoring constraints of the transmembrane domains in the full-length structure. Continuous density was observed for the RAMP1 ECD and transmembrane domain, including the unstructured linker domain, but not for the short C-terminal tail of RAMP (Thr144-Glu-Gly-Ile-Val148), indicating that it is mobile in the active receptor complex (Fig. 1). There was robust density for most of the transmembrane core and loops of CLR, but not for segments of extracellular loop 3 (ECL3) and intracellular loop 3 (ICL3) (Fig. 1, Extended Data Fig. 4). Additional density was observed adjacent to the base of transmembrane domain 2 (TM2) and TM4, which may represent lipid interactions with CLR (Extended Data Fig. 3g). There was a relatively short helix 8 (H8), with no density for the C terminus of CLR beyond $Y402^{8.53}$ (receptor residues in superscript are defined using the class B numbering system[14,22]) or for the far N terminus of the ECD (Fig. 1, Extended Data Fig. 1), indicating that these regions are also mobile. The N terminus of CGRP (A1–V23), which binds within the receptor core, was well-defined in the map, and the majority of side chains in the CGRP C terminus (F27–F37) that interact exclusively with receptor ECDs, were also supported by good density (Extended Data Fig. 4). Similar to salmon calcitonin (sCT) in the G_s-coupled CTR[14], there is a large kink in the peptide that enables interaction across the two receptor domains, with the CGRP linker (K24–N26) being poorly resolved in the map. Within the receptor core, side chains that had limited density were stubbed in the model (Extended Data Fig. 1). There was well-resolved density for the G_s heterotrimer across the receptor interface and between subunits. The α-helical domain of the α-subunit was present only in a small number of the 2D class averages and was masked out during map refinement. In general, the regions of lower resolution or those lacking density were segments of the complex that exhibited higher mobility in molecular dynamics simulations of the full complex (Extended Data Fig. 6, Supplementary Videos 1, 2).

The CLR–RAMP1 interface

The 2D class averages reveal that there is a single predominant orientation of the ECDs of the complex relative to the CLR–RAMP core (Extended Data Fig. 3b). This is in contrast to the variability in ECD orientation observed for the CTR[14]. RAMP1 makes extensive contacts with CLR, with around 23% of its surface being buried within this interface (Fig. 1b, c). The extensive interface across the ECDs has been reported in X-ray crystal structures[12,23]. In contrast to predictions in published models of RAMPs in complex with CLR or CTR[24–27], the RAMP1 transmembrane domain sits at an interface formed by TM3, TM4 and TM5 of CLR, with interactions of the upper half occurring principally with TM5 ($T288^{5.33/ECL2}$, $H289^{5.34/ECL2}$, $I293^{5.38}$) (Fig. 2a, b) and at the base with TM3 ($L231^{3.48}$, $I235^{3.52}$, $T239^{3.56}$, $V243^{ICL2}$) and TM4 ($W254^{4.44}$, $Y255^{4.45}$, $L258^{4.48}$, $F262^{4.52}$) (Fig. 2a, c). These interactions were primarily van der Waals interactions, although there was potential for hydrogen-bond formation between $Y255^{4.45}$ and RAMP1 S141. D113 in the membrane-proximal segment of RAMP1 formed hydrogen bonds with residues in ECL2 proximal to CLR TM4 ($Y278^{ECL2}$) and TM5 ($T288^{5.33}$, $H289^{5.34}$) (Fig. 2b). Alanine mutagenesis studies of CLR residues[28–33] revealed decreased CGRP potency for the $Y278^{ECL2}$, $T288^{5.33/ECL2}$ and $W254^{4.44}$ mutants, with no effect from $H289^{5.34/ECL2}$, $I293^{5.38}$, $T239^{3.56}$, $V243^{ICL2}$ or $Y255^{4.45}$ mutants[32,33], consistent with important but weak interactions between RAMP1 and CLR. Likewise, there was a small decrease in CGRP potency with the RAMP1(D113A) mutant, indicating an indirect effect on CGRP binding[34]. To understand the dynamics of the RAMP1–CLR interface, we performed molecular dynamics simulations, following modelling of missing amino acids and side chains into the full protein complex (Extended Data Fig. 7a, b, Supplementary Table 2, Supplementary Video 1); these simulations confirmed the importance of interactions between RAMP1 D113 and CLR ECL2 (Extended Data Fig. 7a).

© 2018 Springer Nature Limited. All rights reserved.

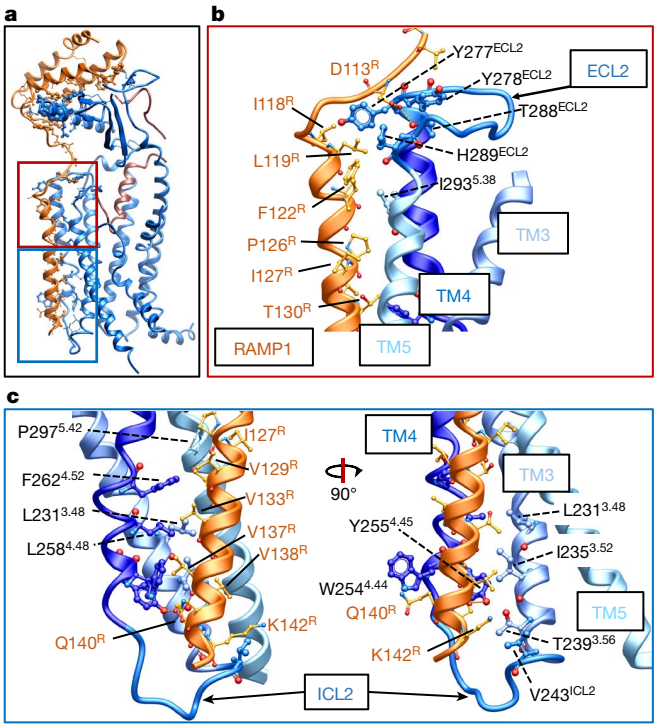


Fig. 2 | RAMP1 forms stable interactions with the CLR core and ECD. **a**, The CGRP–CLR–RAMP1 complex, with the interacting residues depicted in x-stick representation, and the backbone shown in ribbon representation. CGRP, dark red; CLR, blue; RAMP1, dark orange. Boxes indicate regions that are expanded in **b** (red) and **c** (blue). **b**, RAMP1 interacts with ECL2 and the top of TM5 towards the extracellular face of the receptor. **c**, RAMP1 interacts with TM3 and TM4 towards the intracellular face of the receptor; interacting side chains are depicted in x-stick representation and the backbone is depicted in ribbon representation. Superscript R indicates residues of RAMP1.

The simulations also predicted that E47^{ECD} formed persistent hydrogen-bond interactions with RAMP1 R112, in addition to hydrogen bonds to the RAMP1 backbone (G108 and A110) in the linker region. RAMP1 R112 was also predicted to form less frequent hydrogen bonds with D90^{ECD}, but may maintain more persistent ionic interactions. Collectively, these interactions are likely to contribute to the limited mobility of the RAMP1 linker and stable positioning of the ECDs relative to the receptor core (Extended Data Fig. 7a, Supplementary Table 2, Supplementary Videos 2, 3). From the cryo-EM map, there were no resolved interactions between the RAMP and G protein; however, there was no density for the C-terminal tail of RAMP1. Molecular dynamics simulations in which the C terminus of RAMP1 was modelled predicted transient interactions with ICL2 and the αN helix of the Gα protein, with potential interactions that could extend to ICL1 (Supplementary Table 2); nevertheless, this segment was highly mobile in the simulations.

The CGRP-binding site

CGRP forms extensive interactions with the CLR–RAMP1 complex, with 61.5% of its surface buried. Notably, the only direct contact between the peptide and RAMP1 occurs at the far C terminus of the peptide, principally with the cluster of RAMP residues (F83–P85) that have been observed in isolated ECD structures[12] (Fig. 3a). The N-terminal peptide loop that is constrained within CGRP (C2–C7) is deeply buried and extends into an amphipathic α-helix, up to CGRP V23, which forms extensive van der Waals interactions with CLR (Fig. 3d). There are only a small number of hydrogen bonds formed in the static structure between the N terminus of the peptide and the core of CLR; these include interactions between Y292$^{5.37}$ and the backbone of CGRP D3, between H295$^{5.40}$ and CGRP T6, and between S286^{ECL2} and the backbone of CGRP H10 (Fig. 3c, d). Of these, only the interaction between H295$^{5.40}$ and CGRP T6 is functionally important; the H295$^{5.40}$A mutation cause a loss of about 30-fold in CGRP potency on cAMP accumulation[28]. The equivalent residue in CTR is H302$^{5.40}$, which is predicted to form a hydrogen bond with T6 of sCT[14]. Alanine substitution of CGRP T6 leads to a loss of about 80-fold in peptide potency[29], confirming the importance of this bond and other interactions. There are extensive interactions between the peptide and TM3, TM5 and ECL2 of CLR. Below H295$^{5.40}$, a series of amino acids that includes I298$^{5.43}$, L302$^{5.47}$, M223$^{3.40}$ and Y227$^{3.44}$ forms the bottom of the peptide-binding pocket (Fig. 3c, d, Extended Data Fig. 8b). Alanine substitution of CGRP T4 leads to a more than 20-fold reduction in CGRP potency[29]. However, this residue forms only limited interactions with the receptor; side-chain-to-backbone interactions within the peptide that contribute to the loop fold and initiation of the peptide helix may underlie its functional importance. CGRP T9 and H10 pack within an extended cluster of residues that includes T191$^{2.64}$, L195$^{2.68}$, H219$^{3.36}$, S286^{ECL2} and I284^{ECL2} (Fig. 3c, Extended Data Fig. 8b). With the exception of S286^{ECL2}, alanine mutation of these residues caused marked impairment in CGRP signalling[28,30–32] (Extended Data Fig. 8b), with I284^{ECL2} and L195$^{2.68}$ forming a hydrophobic barrier that coincides with the exit of the peptide from the receptor core (Extended Data Fig. 8b); molecular dynamics simulations predict transient hydrogen-bond formation between CGRP T9 and H219$^{3.36}$ (Extended Data Fig. 9e, Supplementary Table 3). Alanine substitution of CGRP T9 causes a 15-fold loss of CGRP potency[29], consistent with the importance of interactions with this side chain. Whereas mutations to amino acids in the distal segment of ECL2 (S286^{ECL2}, D287^{ECL2}, H289^{ECL2}, L291$^{5.36}$) had relatively limited effects on CGRP potency[28] (Extended Data Fig. 8b), ECL2 conformation is critical to CGRP activation of its receptor, with R274$^{4.64}$A and, in particular, W283^{ECL2}A mutations being highly detrimental to CGRP signalling[32] (Extended Data Fig. 8b). These residues are critical for the stable packing of ECL2 in the active structure, similar to those observed in other active, class B GPCR structures[14–16]. There are only limited contacts between ECL1 and the peptide, the most prevalent being from CGRP L16 and S17 to A199^{ECL1}, N200^{ECL1}, Q202^{ECL1} and V205^{ECL1} (Fig. 3b, Extended Data Fig. 8b). Q202^{ECL1} is within weak hydrogen-bond distance of the backbone oxygen of CGRP S17 (Extended Data Fig. 8b); however, alanine mutation of Q202^{ECL1}, N200^{ECL1} or V205^{ECL1} had no effect on CGRP potency, indicating the limited importance of this domain for CGRP activity[30]. CLR and CTR have shorter ECL1 loops compared to the related glucagon receptor[35], or glucagon-like peptide-1 receptor (GLP-1R)[15,16]. These receptors have longer TM2 and TM3 helices (Extended Data Fig. 10a, b) that interact with the extended helix of peptide agonists of these receptors[15,16,35]. In the cryo-EM map, there was no high-resolution density for ECL3, consistent with only limited interaction between CGRP and this receptor segment. This high mobility and lack of persistent interactions was also observed in our molecular dynamics simulations (Extended Data Fig. 9a–f, Supplementary Table 3, Supplementary Video 2), whereas previous alanine mutagenesis studies also support a limited role of this domain in CGRP-mediated cAMP production[28,30] (Extended Data Fig. 8b).

CGRP V8, L12 and L16 are on the same face of the peptide α-helix and sit deep within a groove formed by TM1 and TM7, where they pack among multiple residues on the receptor. Alanine mutations of individual receptor amino acids within this groove have very little effect on CGRP-mediated cAMP production (Extended Data Fig. 8b), consistent with only weak contacts being made by individual receptor amino acids. Nevertheless, alanine substitution of either CGRP L12 or L16 markedly impaired CGRP potency[36], indicating that the packing of the hydrophobic face of the peptide helix is critical for receptor activation.

In the cryo-EM structure, CGRP R11 forms polar interactions with the backbone of CGRP T4 and C2, with potential salt-bridge interactions with CGRP D3 and D366$^{7.39}$ on the receptor, and may contribute to stability of the CGRP loop conformation (Fig. 3C).

© 2018 Springer Nature Limited. All rights reserved.

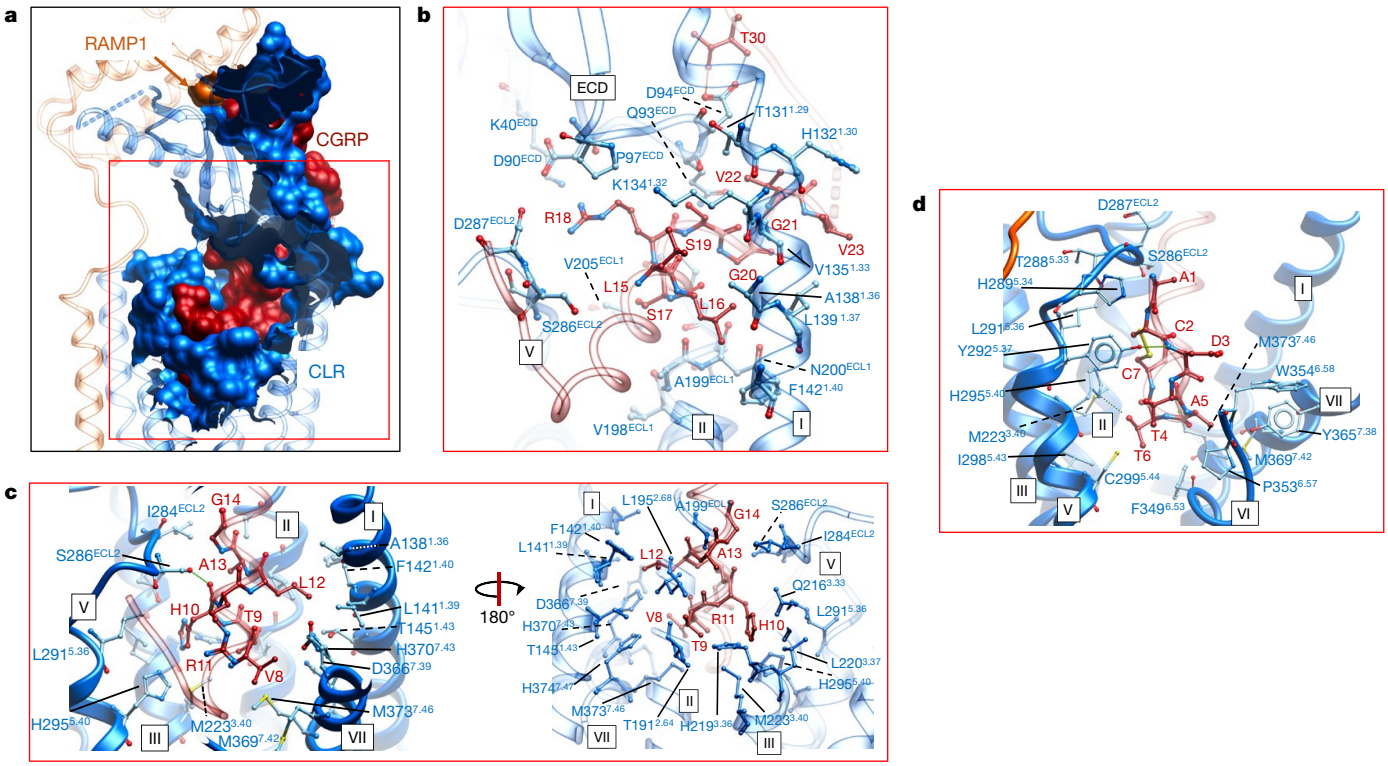


Fig. 3 | The CGRP-binding site. a, The CGRP-interaction surfaces (amino acids within 5 Å) of CLR (blue) or RAMP1 (dark orange), illustrating how the peptide N terminus is buried within CLR. CGRP is shown in dark red surface representation. **b**–**d**, Amino acid side chains of CLR proximal to CGRP residues; amino acids are shown in x-stick representation with carbons in blue (CLR) or dark red (CGRP), and other atoms are coloured by type. Receptor transmembrane helices are numbered using roman numerals. **b**, Contact residues in CGRP, L15–V23; T30, which forms two hydrogen bonds with CLR D94^{ECD} is also shown. **c**, CGRP contact residues V8–G14. **d**, CGRP contact residues A1–C7. There are very few hydrogen bonds between the peptide N terminus and CLR in the static structure.

In molecular dynamics simulations, CGRP R11 formed persistent hydrogen bonds with D366$^{7.39}$, although these interactions are not observed in the cryo-EM map. CGRP R18 is within salt-bridge distance of D287^{ECL2} and D90^{ECD}, and forms a hydrogen bond with D287^{ECL2} in nearly 25% of frames in the simulation (Fig. 3b, Extended Data Fig. 9d, e).

The resolutions of the peptide C terminus and receptor ECDs are lower than the resolution in the receptor core, and they were primarily modelled by rigid-body fitting of the available X-ray structure (PDB: 4RWG[12]). To test the stability of interactions in the fully active structure, we ran 6.4-μs molecular dynamics simulations. Our data are consistent with the interactions previously reported in the isolated ECD structure[12], and are summarized in Extended Data Fig. 9a–f and Supplementary Table 3. The main intermolecular interactions were between CGRP T30 and D94^{ECD} (Fig. 3b, Extended Data Fig. 9e), and between the amide of CGRP F37 and backbone atoms of T122^{ECD} (Extended Data Fig. 9e). There were no persistent hydrogen bonds between CGRP and RAMP1. The critical importance of interactions between the C terminus of CGRP (F27–F37) and the ECDs of CLR and RAMP1 for CGRP signalling has been highlighted by previous mutagenesis studies[12,23,34,37,38], and are illustrated in Extended Data Fig. 8a. The extent to which this is dependent upon the stability of the relative positioning of the ECD to the receptor core is unclear, but RAMP1 is a major contributor to the limited conformational flexibility of the ECD domain of CLR (Supplementary Video 3).

Comparisons with the CTR structure

CTR is most closely related to CLR, and can also interact with RAMP1 to form a high-affinity CGRP receptor[1]. Therefore, we compared the structure of the sCT–CTR–G_s complex to the CGRP–receptor complex. Owing to the relatively limited resolution in the peptide-binding domain and N terminus of the CTR, comparisons were limited to the backbone structures in these regions. Overall, there was a high degree of similarity between the CLR and CTR structures, with both exhibiting an extended TM1 α-helical stalk that interconnects the receptor core and ECD, and a similar organization of the upper segments of TM6 and TM7 to accommodate the bulk of the cysteine-bridged loops of the peptides (Fig. 4a).

The largest difference between the CTR and CLR structures was in the orientation of the ECD relative to the receptor core (Fig. 4a, b). This located the C terminus of the peptides at nearly equivalent positions, with the N-terminal activation domain of the peptides also occupying a similar binding cavity (Fig. 4b). Within the receptor core, there was an inward shift of approximately 2 Å of the CLR relative to CTR at the apex of TM5, which is likely to be a result of the interaction of RAMP1 with this domain (Fig. 4c). There is a high degree of sequence conservation between CLR and CTR among the residues that contacted the RAMP (Fig. 4d), which may explain the similar broad specificity for RAMP interaction of these receptors. In previous simulations of the CTR bound to human CT versus sCT, there was destabilization of ECL2 for human CT relative to the sCT-bound receptor that was indicative of a role for conformational dynamics of this receptor domain in ligand interaction and efficacy[18]. The interactions of RAMP1 with ECL2 may therefore contribute to peptide selectivity and/or efficacy.

At the base of the receptor, the structured H8 of CLR was much shorter than that of CTR (Fig. 4a), and consequently exhibited more-limited interaction with the Gβ-subunit. Nevertheless, truncation studies of the CTR C terminus indicated that only the segment that is also present in the CGRP receptor structure is functionally important for G_s-mediated signalling[14]. Perhaps more relevant, although the α-helix 5 of the Ras-like domain of $G\alpha_s$ ($G\alpha_s$-Ras) is aligned between the two structures, there are differences within the G protein, particularly with respect to the positioning of the αN helix of $G\alpha_s$-Ras; these differences are propagated across the β- and γ-subunits (Fig. 4e).

Broader comparison of protein interactions of G_s to include the structures of GLP-1R bound to either exendin-P5[15] or GLP-1[16]

© 2018 Springer Nature Limited. All rights reserved.

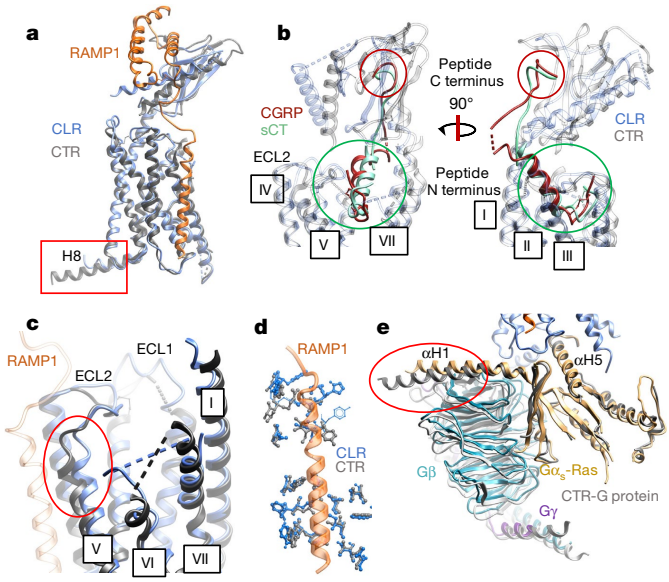


Fig. 4 | The CTR and CGRP receptor complexes display similar backbone conformations but have distinct conformations of the $G\alpha_s$-Ras domain. a, Alignment of the CLR–RAMP1–CTR (blue, dark orange and grey ribbon, respectively) structures; for the CTR the ECD is from the X-ray structure of the sCT–CTR X-ray structure (PDB: 5II0[39]), following rigid-body fitting to the CTR cryo-EM map[14]. Receptor transmembrane helices are numbered using roman numerals. **b**, Magnified view of the peptide-binding sites. CGRP (dark red) and sCT (green) are shown as ribbon, CLR (blue) and CTR (grey) are shown as transparent ribbon. The circles highlight the similarities in position of the peptide N- (green) and C- (red) termini. RAMP1 has been omitted for clarity. **c**, Magnified view highlighting differences in the upper segment of TM5 (red circle). **d**, Overlap in RAMP1-contact residues between CLR (blue x-stick) and CTR (grey x-stick). **e**, The $G\alpha_s$–Ras–H5 is superimposed in the two structures, but the α-H1 helix is in a different orientation (red circle) and leads to distinct positioning of the Gβ- and Gγ-subunits. The CTR G protein is shown as grey ribbon, the CGRP receptor G protein as coloured ribbon. $G\alpha_s$-Ras, gold; Gβ, cyan; Gγ, dark purple. Regions of the receptor structures that are missing in the PDB files are shown as dashed lines.

(Extended Data Fig. 10a, b) also revealed differences in the relative positioning of G_s. However, this was principally due to translational differences in the engagement of the receptors and α-helix 5 (Extended Data Fig. 10c), with strong overlap in the backbone of the Gα-subunit when these are aligned (Extended Data Fig. 10d). ICL2 of CLR and of CTR are longer than that of GLP-1R, and there is an increase of about 2 Å in the outward movement of the base of TM6 of CLR and CTR compared to GLP-1R (Extended Data Fig. 10a); these differences are likely to account for the translational differences in engagement of the G_s protein by GLP-1R.

Stability of the complex in the absence of RAMP1

In molecular dynamics simulations of the complex in the presence and absence of RAMP1, the orientation of the ECD of CLR remains relatively stable in CLR–CGRP–RAMP1–Gαβγ–Nb35 (6.4-μs simulation) and CLR–CGRP–RAMP1–mini-Gα (2-μs simulation), but not in CLR–CGRP–mini-Gα (2-μs simulation) (Supplementary Videos 1–3). In the absence of RAMP1, only CGRP and TM1—with its extension—hold the ECD in place relative to the transmembrane domain. The N-terminal region (A1–R18) of CGRP is stable (Supplementary Video 2) even in the absence of RAMP1, but the C-terminal region is affected by the mobility of the CLR ECD and is much more mobile in the absence of RAMP1 (Extended Data Fig. 6, Supplementary Video 3). A consequence of this C-terminal mobility in the absence of RAMP1 is reduced persistence of hydrogen bonds formed by CGRP in this region (Supplementary Table 4).

RAMP1 provides additional stability to ECL2, a major contact point for CGRP—though this loop is relatively stable even in the absence of RAMP1. In the simulation in the absence of RAMP1, there is a marked reduction in the persistence of hydrogen bonds between $R274^{4.64}$ and $D280^{ECL2}$ (Supplementary Table 5). In the cryo-EM structure, these two residues form a salt-bridge interaction, and this interaction in the presence of RAMP1 is likely to affect signal propagation. Indeed, mutagenesis of either of these residues greatly affects CGRP-mediated cAMP signalling[28,32]. RAMP1 interaction does not affect the mobility of the distally located ECL1 and ICL3. The least mobile points of each transmembrane domain generally correspond to points of helix intersection; for TM3 this is in the vicinity of $Y227^{3.44}$, which provides a deep stable contact point for CGRP.

Whereas these simulations provide insight into the contribution of RAMP1 to the preformed active complex, this complex does not form in the absence of RAMP1, even where CLR is present at the cell surface[6], indicating that the CLR–RAMP1 interaction is also critical for initial peptide binding and presentation to the receptor core.

In conclusion, the structure of the CGRP–CLR–RAMP1–G_s complex provides insight into the organization of functionally important heteromeric GPCR complexes. RAMP1 causes marked stabilization of the ECD of CLR, and therefore has a critical role in ligand presentation to the receptor core. It further enhances stability of the transmembrane domain interface and ECL2, which are important for propagation of peptide-induced signalling. This study provides a framework for the development of novel therapeutics that target the CGRP system.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0535-y.

Received: 1 June 2018; Accepted: 26 June 2018;
Published online 12 September 2018.

1. Hay, D. L., Garelja, M. L., Poyner, D. R. & Walker, C. S. Update on the pharmacology of calcitonin/CGRP family of peptides: IUPHAR Review 25. *Br. J. Pharmacol.* **175**, 3–17 (2018).
2. Russell, F. A., King, R., Smillie, S. J., Kodji, X. & Brain, S. D. Calcitonin gene-related peptide: physiology and pathophysiology. *Physiol. Rev.* **94**, 1099–1142 (2014).
3. Dickerson, I. M., Bussey-Gaborski, R., Holt, J. C., Jordan, P. M. & Luebke, A. E. Maturation of suprathreshold auditory nerve activity involves cochlear CGRP-receptor complex formation. *Physiol. Rep.* **4**, e12869 (2016).
4. Walker, C. S. et al. Mice lacking the neuropeptide α-calcitonin gene-related peptide are protected against diet-induced obesity. *Endocrinology* **151**, 4257–4269 (2010).
5. Karsan, N. & Goadsby, P. J. Calcitonin gene-related peptide and migraine. *Curr. Opin. Neurol.* **28**, 250–254 (2015).
6. McLatchie, L. M. et al. RAMPs regulate the transport and ligand specificity of the calcitonin-receptor-like receptor. *Nature* **393**, 333–339 (1998).
7. Barbash, S., Lorenzen, E., Persson, T., Huber, T. & Sakmar, T. P. GPCRs coevolved with receptor activity-modifying proteins, RAMPs. *Proc. Natl Acad. Sci. USA* **114**, 12015–12020 (2017).
8. Hay, D. L. & Pioszak, A. A. Receptor activity-modifying proteins (RAMPs): new insights and roles. *Annu. Rev. Pharmacol. Toxicol.* **56**, 469–487 (2016).
9. Routledge, S. J., Ladds, G. & Poyner, D. R. The effects of RAMPs upon cell signalling. *Mol. Cell. Endocrinol.* **449**, 12–20 (2017).
10. Wootten, D. et al. Receptor activity modifying proteins (RAMPs) interact with the VPAC2 receptor and CRF1 receptors and modulate their function. *Br. J. Pharmacol.* **168**, 822–834 (2013).
11. Christopoulos, A. et al. Novel receptor partners and function of receptor activity-modifying proteins. *J. Biol. Chem.* **278**, 3293–3297 (2003).
12. Booe, J. M., Warner, M. L., Roehrkasse, A. M. & Pioszak, A. A. Structural basis for receptor activity-modifying protein-dependent selective peptide recognition by a G protein-coupled receptor. *Mol. Cell* **58**, 1040–1052 (2015).
13. Booe, J. M. et al. Probing the mechanism of receptor activity-modifying protein modulation of GPCR ligand selectivity through rational design of potent adrenomedullin and calcitonin gene-related peptide antagonists. *Mol. Pharmacol.* **93**, 355–367 (2018).
14. Liang, Y. L. et al. Phase-plate cryo-EM structure of a class B GPCR-G-protein complex. *Nature* **546**, 118–123 (2017).
15. Liang, Y. L. et al. Phase-plate cryo-EM structure of a biased agonist-bound human GLP-1 receptor-G_s complex. *Nature* **555**, 121–125 (2018).
16. Zhang, Y. et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **546**, 248–253 (2017).

© 2018 Springer Nature Limited. All rights reserved.

17. Wootten, D. et al. The extracellular surface of the GLP-1 receptor is a molecular trigger for biased agonism. *Cell* **165**, 1632–1643 (2016).
18. Dal Maso, E. et al. Extracellular loops 2 and 3 of the calcitonin receptor selectively modify agonist binding and efficacy. *Biochem. Pharmacol*. **150**, 214–244 (2018).
19. Liang, Y.-L. et al. Dominant negative G proteins enhance formation and purification of agonist-GPCR–G protein complexes for structure determination *ACS Pharmacol Transl Sci* https://doi.org/10.1021/acsptsci.8b00017 (2018).
20. Khoshouei, M. et al. Volta phase plate cryo-EM of the small protein complex Prx3. *Nat. Commun*. **7**, 10534 (2016).
21. Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat. Commun*. **8**, 16099 (2017).
22. Wootten, D., Simms, J., Miller, L. J., Christopoulos, A. & Sexton, P. M. Polar transmembrane interactions drive formation of ligand-specific and signal pathway-biased family B G protein-coupled receptor conformations. *Proc. Natl Acad. Sci. USA* **110**, 5211–5216 (2013).
23. Ter Haar, E. et al. Crystal structure of the ectodomain complex of the CGRP receptor, a class-B GPCR, reveals the site of drug antagonism. *Structure* **18**, 1083–1093 (2010).
24. Watkins, H. A. et al. Receptor activity-modifying proteins 2 and 3 generate adrenomedullin receptor subtypes with distinct molecular properties. *J. Biol. Chem*. **291**, 11657–11675 (2016).
25. Woolley, M. J. et al. Receptor activity-modifying protein dependent and independent activation mechanisms in the coupling of calcitonin gene-related peptide and adrenomedullin receptors to Gs. *Biochem. Pharmacol*. **142**, 96–110 (2017).
26. Weston, C. et al. Receptor activity-modifying protein-directed G protein signaling specificity for the calcitonin gene-related peptide family of receptors. *J. Biol. Chem*. **291**, 21925–21944 (2016).
27. Bower, R. L. et al. Molecular signature for receptor engagement in the metabolic peptide hormone amylin. *ACS Pharmacol. Transl. Sci*. https://doi.org/10.1021/acsptsci.8b00002 (2018).
28. Woolley, M. J. et al. Understanding the molecular functions of the second extracellular loop (ECL2) of the calcitonin gene-related peptide (CGRP) receptor using a comprehensive mutagenesis approach. *Mol. Cell. Endocrinol*. **454**, 39–49 (2017).
29. Hay, D. L. et al. Structure-activity relationships of the N-terminus of calcitonin gene-related peptide: key roles of alanine-5 and threonine-6 in receptor activation. *Br. J. Pharmacol*. **171**, 415–426 (2014).
30. Barwell, J., Conner, A. & Poyner, D. R. Extracellular loops 1 and 3 and their associated transmembrane regions of the calcitonin receptor-like receptor are needed for CGRP receptor function. *Biochim. Biophys. Acta* **1813**, 1906–1916 (2011).
31. Vohra, S. et al. Similarity between class A and class B G-protein-coupled receptors exemplified through calcitonin gene-related peptide receptor modelling and mutagenesis studies. *J. R. Soc. Interface* **10**, 20120846 (2012).
32. Woolley, M. J. et al. The role of ECL2 in CGRP receptor activation: a combined modelling and experimental approach. *J. R. Soc. Interface* **10**, 20130589 (2013).
33. Conner, A. C., Simms, J., Howitt, S. G., Wheatley, M. & Poyner, D. R. The second intracellular loop of the calcitonin gene-related peptide receptor provides molecular determinants for signal transduction and cell surface expression. *J. Biol. Chem*. **281**, 1644–1651 (2006).
34. Simms, J. et al. Structure-function analysis of RAMP1 by alanine mutagenesis. *Biochemistry* **48**, 198–205 (2009).
35. Zhang, H. et al. Structure of the glucagon receptor in complex with a glucagon analogue. *Nature* **553**, 106–110 (2018).
36. Simms, J. et al. Photoaffinity cross-linking and unnatural amino acid mutagenesis reveal insights into calcitonin gene-related peptide binding to the calcitonin receptor-like receptor/receptor activity-modifying protein 1 (CLR/RAMP1) complex. *Biochemistry* **57**, 4915–4922 (2018).
37. Garelja, M. L. et al. Receptor activity modifying proteins have limited effects on the Class B G protein-coupled receptor calcitonin receptor-like receptor stalk. *Biochemistry* **57**, 1410–1422 (2018).
38. Hay, D. L., Christopoulos, G., Christopoulos, A. & Sexton, P. M. Determinants of 1-piperidinecarboxamide, *N*-[2-[[5-amino-*l*-[[4-(4-pyridinyl)-*l*-piperazinyl] carbonyl]pentyl]amino]-1-[(3,5-dibromo-4-hydroxyphenyl)methyl]-2-oxoethyl]-4-(1,4-dihydro-2-oxo-3(2*H*)-quinazolinyl) (BIBN4096BS) affinity for calcitonin gene-related peptide and amylin receptors–the role of receptor activity modifying protein 1. *Mol. Pharmacol*. **70**, 1984–1991 (2006).
39. Johansson, E. et al. Type II turn of receptor-bound salmon calcitonin revealed by X-ray crystallography. *J. Biol. Chem*. **291**, 13689–13698 (2016).

Acknowledgements The work was supported by the Monash University Ramaciotti Centre for Cryo-Electron Microscopy, National Health and Medical Research Council of Australia (NHMRC) project grant (1120919), and NHMRC program grant (1055134). P.M.S. and A.C. are NHMRC Principal and Senior Principal Research Fellows, respectively. D.W. is a NHMRC Career Development Fellow, and C.K. is a NHMRC CJ Martin Fellow. A.G. is an Australian Research Council DECRA Fellow. D.L.H. is a James Cook Research Fellow and is supported by the Marsden Fund (both Royal Society of New Zealand). C.A.R. is a Royal Society Industry Fellow and acknowledges support from the BBSRC (BB/M006883/1). We are grateful to G. Christopoulos and V. Julita for assay and technical support, T. Coudrat for initial homology modelling of CLR from the active CTR, and to S. Furness, P. Zhao and D. Thal for useful discussion.

Reviewer information *Nature* thanks K. Caron, S, Dang, B. Wu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.-L.L performed virus production, insect cell expression, purification, negative stain electron microscopy, data acquisition and analysis, and prepared samples for cryo-EM, was responsible for model building and refinement, and assisted with manuscript preparation. M.K. performed cryo-sample preparation, phase-plate imaging, data collection, electron microscopy data processing and analysis, calculated the cryo-EM map and assisted with manuscript preparation. G.D. performed molecular dynamics simulations and assisted in manuscript preparation. A.G. assisted with model building and refinement and contributed to manuscript preparation. T.S.P. assisted with model building and refinement and reviewed the manuscript. C.K. performed cell-based assays and data analysis and reviewed the manuscript. M.R. performed preliminary cryo-EM screening and reviewed the manuscript. J.M.P. and W.B. organized and managed the Volta phase-plate development project. D.L.H. provided insights into the CGRP receptor, assisted with data interpretation, and reviewed the manuscript. L.J.M. provided insights into class B GPCRs, assisted with data interpretation and reviewed the manuscript. A.C. assisted with data interpretation and manuscript preparation. C.A.R. designed molecular dynamics simulations, assisted in data interpretation and contributed to writing of the manuscript. D.W. was responsible for overall project strategy and management, data analysis and interpretation and contributed to writing of the manuscript. P.M.S. was responsible for overall project strategy and management, data interpretation and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0535-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0535-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.W. or P.M.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Constructs. CLR was modified to include an N-terminal Flag-tag epitope and a C-terminal 8× His tag; both tags are removable by 3C protease cleavage. The construct was generated in both mammalian and insect cell expression vectors. RAMP1 was modified to include an N-terminal Flag-tag epitope. For both constructs, the natural signal peptide was replaced with that of haemagglutinin to improve expression (Extended Data Fig. 1).

Expression in insect cells. CLR, RAMP1, DNGα_s[15], His_6-tagged human Gβ1 and Gγ2 were expressed in *Trichoplusia ni* insect cells (Expression Systems) using baculovirus. Cell cultures were grown in ESF 921 serum-free medium (Expression Systems) to a density of 4×10^6 cells per ml and then infected with the three baculoviruses CLR, RAMP1, DNG$_{\alpha s}$ and Gβ1γ2 at a ratio of 1:5:2:1. Culture was collected by centrifugation 48 h after infection and cell pellet was stored at −80 °C.

Complex purification. Cell pellet was thawed in 20 mM HEPES pH 7.4, 50 mM NaCl, 2 mM $MgCl_2$ supplemented with cOmplete Protease Inhibitor Cocktail tablets (Roche). Complex formation was initiated by addition of 10 μM human αCGRP (Chinapeptide), Nb35–His (10 μg/ml), 3C protease (10 μg/ml) and apyrase (25 mU/ml, NEB); the suspension was incubated for 1 h at room temperature. Membranes were collected by centrifugation at 30,000*g* for 30 min. Complexes from membranes were solubilized by 0.5% (w/v) lauryl maltose neopentyl glycol (LMNG, Anatrace) supplemented with 0.03% (w/v) cholesteryl hemisuccinate (CHS, Anatrace) for 2 h at 4 °C in the presence of 1 μM CGRP and apyrase (25 mU/ml, NEB). Insoluble material was removed by centrifugation at 30,000*g* for 30 min and the solubilized complex was immobilized by batch binding to Ni-NTA resin. The resin was packed into a glass column and washed with 20 column volumes of 20 mM HEPES pH 7.4, 100 mM NaCl, 2 mM $MgCl_2$, 0.01% (w/v) LMNG and 0.006% (w/v) CHS, 1 μM CGRP, before bound material was eluted in buffer containing 250 mM imidazole. The Ni-NTA-purified fraction was immobilized by batch binding to M1 anti-Flag affinity resin in the presence of 3 mM $CaCl_2$. The resin was packed into a glass column and washed with 20 column volumes of 20 mM HEPES pH 7.4, 100 mM NaCl, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 1 μM CGRP, 0.01% (w/v) LMNG and 0.006% (w/v) CHS before bound material was eluted in buffer containing 5 mM EGTA and 0.1 mg/ml Flag peptide. The complex was then concentrated using an Amicon Ultra Centrifugal Filter (MWCO 100 kDa) and subjected to size-exclusion chromatography on a Superose 6 Increase 10/300 column (GE Healthcare) that was pre-equilibrated with 20 mM HEPES pH 7.4, 100 mM NaCl, 2 mM $MgCl_2$, 1 μM CGRP, 0.01% (w/v) LMNG and 0.006% (w/v) CHS. Eluted fractions consisting of receptor and G-protein complex were pooled and concentrated. Final yield of purified complex was approximately 0.3 mg per litre of insect cell culture.

SDS–PAGE and western blot analysis. Sample collected from size-exclusion chromatography was analysed by SDS–PAGE and western blot as previously described[15]. For SDS–PAGE, precast gradient TGX gels (Bio-Rad) were used. The final elution peak from size-exclusion chromatography was stained using Instant Blue (Expedeon).

Modelling into cryo-EM density. An initial template for CLR was generated by homology modelling using the cryo-EM structure of human CTR (PDB: 5UZ7)[14], performed with the Molsoft ICM modelling software[40]. Manual adjustment and rebuilding was performed in Coot[41]. Owing to limited density in CLR and RAMP1 ECD regions, we used the high-resolution X-ray crystal structure (PDB: 4RWG)[12] for modelling. ECDs of CLR and RAMP1 were, separately, rigid-body fitted into density before the final iteration of global refinement. DNGα$_s$, Gβ1, Gγ2 and Nb35 models were taken from the GLP1-R–G$_s$–ExP5 structure (PDB: 6B3J)[15]. The CGRP peptide and RAMP1 transmembrane domain were modelled manually. The final model was subjected to global refinement and minimization in real space using the module 'phenix.real_space_refine' in PHENIX[42]. Validation was performed in MolProbity[43].

Preparation of vitrified specimen. Electron microscopy grids (Quantifoil, 200-mesh copper R1.2/1.3) were glow-discharged for 30 s using Harrick plasma cleaner (Harrick). Four microlitres of sample was applied on the grid in the Vitrobot Mark IV chamber (Thermo Fisher Scientific). The chamber of the Vitrobot was set to 100% humidity at 4 °C. The sample was blotted for 4.5 s with a blot force of 20 and then plunged into propane–ethane mixture (37% ethane and 63% propane).

Data acquisition. Datasets were collected on a Thermo Fisher Scientific Titan Krios microscope operated at 300 kV (FEI) equipped with a Gatan Quantum energy filter, a Gatan K2 summit direct electron camera (Gatan) and a Volta phase plate (Thermo Fisher Scientific). Movies were taken in EFTEM nanoprobe mode, with 50-μm C2 aperture, at a calibrated magnification of 47170 corresponding to a magnified pixel size of 1.06 Å. Each movie comprises 50 subframes with a total dose of 50 e^- per Å^2, exposure time was 13 s with a dose rate of 4.8 e^- pixel^{-1} s^{-1} on the detector. Data acquisition was done using SerialEM software at −600-nm defocus[44].

Data processing. A total of 3,180 movies were collected and subjected to motion correction using MotionCor2[45]. CTF estimation was done using Gctf software[46] on non-dose-weighted micrographs. The particles were picked using Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/). An initial model was made using the common-line approach in EMAN2[47], based on a few automatically picked micrographs and using the common-line approach. The particles were extracted in RELION v.2.01b1[48] using a box size of 200 pixels. A total of 1,205,000 picked particles were subjected to 2D classification with 100 classes, followed by 3D classification. After selecting the best-looking class, with 407,000 particles, 3D auto-refinement was performed in RELION v.2.01b1. The final map was sharpened with a B-factor of −50 Å^2. The processing workflow is outlined in Extended Data Fig. 3c. Model overfitting was evaluated by randomly displacing all atoms by 0.5 Å and refined against one cryo-EM half map. Fourier shell correlation curves were calculated between the resulting model and the half map used for refinement, the resulting model and the other half map for cross validation, and the final refined model and the full map (Extended Data Fig. 3f).

Mammalian cell cAMP assays. COS-7 cells, which were confirmed to be free from mycoplasma, were transfected in suspension in 96-well plates (10,000 cells per well) with 50 ng CLR + 50 ng human RAMP1 using 600 ng polyethylenimine (PEI). The transfection was performed in DMEM with 5% FBS, 200 μl total volume per well, and cells were incubated for 48 h at 37 °C, 5% CO_2. cAMP detection was performed as described[49]. All values were converted to cAMP concentration using a cAMP standard curve performed in parallel and data were subsequently normalized to the response to 100 μM forskolin.

Conformational clustering of CGRP ECL3 and RAMP1. The missing loops throughout CLR were generated using PLOP[50], which has been shown to be effective in generating GPCR loop conformations[51]. The missing side chains were iteratively optimized to convergence using PLOP. In addition, to enhance the conformational sampling of ECL3, which is likely to interact with the CGRP peptide, a preliminary clustering of 4,000 different loop models generated using Modeller 9.16[52] was performed by means of the Clustering VMD plugin (available at http://physiology.med.cornell.edu/ faculty/hweinstein/vmdplugins/clustering/). Conformational clustering was based on the coordinates of side chains belonging to residues $W354^{6.58}$, $R355^{6.59}$, $P356^{ECL3}$, $E357^{ECL3}$, $K359^{ECL3}$, $I360^{ECL3}$, $A361^{ECL3}$ and $E362^{ECL3}$. A total of 10 clusters was generated with a root mean standard deviation cut-off value of 3 Å and a representative structure with a low distributed optimized potential energy score from the four most-populated ensembles was extracted and prepared for molecular dynamics simulations.

A similar approach was employed for clustering the modelled RAMP1 C terminus (residues T144–V148): the original PLOP-generated conformation was combined with each of the 4 initially selected ECL3 conformations, whereas a highly distinct RAMP1 C-terminus orientation was arbitrarily combined with ECL3 conformation number 1.

Molecular dynamics simulations. A total of seven systems was prepared for molecular dynamics simulations with the CHARMM36 force field[53] (Supplementary Table 6) using a multistep procedure that combines Python HTMD[54] and tool command language scripts. Hydrogen atoms were first added by means of pdb2pqr[55] and propka[56] software (considering a simulated pH of 7.0); the protonation of titratable side chains was checked by visual inspection. CLR and RAMP1 were embedded in a square 116 Å × 116 Å 1-palmitoyl-2-oleyl-*sn*-glycerol-3-phospho-choline (POPC) bilayer (previously built using the VMD Membrane Builder plugin 1.1: http://www.ks.uiuc.edu/Research/vmd/plugins/membrane/) through an insertion method[57]. More precisely, the opportune receptor orientation was obtained by superposing CLR coordinates on the CTR structure retrieved from the OPM database[58]. Lipids overlapping the receptor transmembrane-domain bundle and RAMP1 were removed and TIP3P water molecules[59] were added to the simulation box (116 Å × 116 Å × 185 Å) by means of the VMD Solvate plugin v.1.5 (http://www.ks.uiuc.edu/Research/vmd/plugins/solvate/). Overall charge neutrality was finally reached by adding Na^+/Cl^- counter ions (final ionic strength of 0.150 M), using the VMD Autoionize plugin 1.3 (http://www.ks.uiuc.edu/Research/vmd/plugins/autoionize/).

To evaluate the influence exerted by RAMP1 on the CGRP–CLR complex, two simplified systems were embedded in 96 Å × 96 Å POPC bilayers and solvated as described above: one was composed of CLR–CGRP–RAMP1 and the C terminus (residues $N371^G$–$L394^G$) of the G-protein α-subunit (CLR–CGRP–RAMP1–Gα(371–394)); the other was formed by CLR–CGRP and the C terminus of the Gα-subunit (CLR–CGRP–Gα(371–394)). The original PLOP-generated conformations of CLR and RAMP1 were used.

Systems equilibration and molecular dynamics settings. The molecular dynamics engine ACEMD[60] was used for both the equilibration and productive simulations. Equilibration was achieved in isothermal-isobaric conditions (NPT) using the Berendsen barostat[61] (target pressure 1 atm) and the Langevin thermostat[62]

© 2018 Springer Nature Limited. All rights reserved.

(target temperature 300 K) with a low damping of 1 ps^{-1}. A three-stage procedure was performed (integration time-step of 2 fs): first, clashes between protein and lipid atoms were reduced through 2,500 conjugate-gradient minimization steps, then a 2-ns-long molecular dynamics simulation was run with a positional constraint of 1 kcal mol^{-1} $Å^{-2}$ on protein and lipid phosphorus atoms. During the second stage, 40 ns of simulation was performed, constraining only the protein atoms, whereas in the last equilibration stage, positional constraints were applied only to the protein backbone alpha carbons, for a further 5 ns.

Supplementary Table 6 summarizes all the simulations performed. Trajectories were computed with an integration time-step of 4 fs in the canonical ensemble (NVT) at 300 K, using a thermostat damping of 0.1 ps^{-1} and the M-SHAKE algorithm[63] to constrain the bond lengths involving hydrogen atoms. The cut-off distance for electrostatic interactions was set at 9 Å, with a switching function applied beyond 7.5 Å. Long-range Coulomb interactions were handled using the particle mesh Ewald summation method (PME)[64] by setting the mesh spacing to 1.0 Å.

Molecular dynamics analysis. Atomic contacts, hydrogen bonds and root mean square fluctuation (RMSF) were computed using VMD[65]. A contact was considered productive if the distance between two atoms was smaller than 3.5 Å. For hydrogen-bond detection, a donor-acceptor distance of 3 Å and an angle value of 160° were set as geometrical cut-offs. The hydrogen-bond persistence is defined as the number of frames in which the hydrogen bond is formed, divided by the total number of frames × 100. The RAMP1 influence on van der Waals contacts and hydrogen bonds was evaluated by computing the numerical difference between the total numbers of contacts or hydrogen bonds between each CLR and CGRP side chain during the simulations in the presence and absence of RAMP1.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All relevant data are available from the authors and/or included in the manuscript or Supplementary Information. Atomic coordinates and the cryo-EM density map have been deposited in the Protein Data Bank under accession number 6E3Y and the Electron Microscopy Data Bank, entry EMD-8978.

40. Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002 (1994).
41. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
42. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
44. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 3651 (2005).
45. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
46. Zheng, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
47. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
48. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
49. Hager, M. V., Clydesdale, L., Gellman, S. H., Sexton, P. M. & Wootten, D. Characterization of signal bias at the GLP-1 receptor induced by backbone modification of GLP-1. *Biochem. Pharmacol.* **136**, 99–108 (2017).
50. Jacobson, M. P. et al. A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351–367 (2004).
51. Goldfeld, D. A., Zhu, K., Beuming, T. & Friesner, R. A. Successful prediction of the intra- and extracellular loops of four G-protein-coupled receptors. *Proc. Natl Acad. Sci. USA* **108**, 8275–8280 (2011).
52. Eswar, N. et al. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* **15**, 5.6.1– 5.6.30 (2006).
53. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
54. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: high-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
55. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–667 (2004).
56. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical pK predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
57. Sommer, B. Membrane packing problems: a short review on computational membrane modeling methods and tools. *Comput. Struct. Biotechnol. J.* **5**, e201302014 (2013).
58. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **22**, 623–625 (2006).
59. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).
60. Harvey, M. J., Giupponi, G. & Fabritiis, G. D. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
61. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684 (1984).
62. Loncharich, R. J., Brooks, B. R. & Pastor, R. W. Langevin dynamics of peptides: the frictional dependence of isomerization rates of *N*-acetylalanyl-*N*′-methylamide. *Biopolymers* **32**, 523–535 (1992).
63. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**, 501–508 (2001).
64. Essmann, U. et al. A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577 (1995).
65. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

© 2018 Springer Nature Limited. All rights reserved.

Native protein sequence
HA signal peptide
FLAG epitope
3C cleavage site
His tag
TMs
Omitted residues due to lack of EM density
Stubbed residues

CLR – Expression construct

```
  1 MEKKCTLYFL VLLPFFMILV T
  1 MKTIIALSYI FCLVFADYKD DDDLEVLFQG PAELEESPED SIQLGVTRNK

 51 IMTAQYECYQ KIMQDPIQQA EGVYCNRTWD GWLCWNDVAA GTESMQLCPD

101 YFQDFDPSEK VTKICDQDGN WFRHPASNRT WTNYTQCNVN THEKVKTALN
                  stalk, TM1                          TM2
151 LFYLTIIGHG LSIASLLISL GIFFYFKSLS CQRITLHKNL FFSFVCNSVV
                                          TM3
201 TIIHLTAVAN NQALVATNPV SCKVSQFIHL YLMGCNYFWM LCEGIYLHTI
                          TM4
251 IVVAVFAEKQ HLMWYYFLGW GFPLIPACIH AIARSLYYND NCWISSDTHL
                  TM5                                  TM6
301 LYIIHGPICA ALLVNLFFLL NIVRVLITKL KVTHQAESNL YMKAVRATLI
                                      TM7
351 LVPLLGIEFV LIPWRPEGKI AEEVYDYIMH ILMHFQGLLV STIFCFFNGE
      H8
401 VQAILRRNWN QYKIQFGNSF SNSEALRSAS YTVSTISDGP GYSHDCPSEH

451 LNGKSIHDIE NVLLKPENLY NPAGLEVLFQ GPHHHHHHHH
```

RAMP1- Expression construct

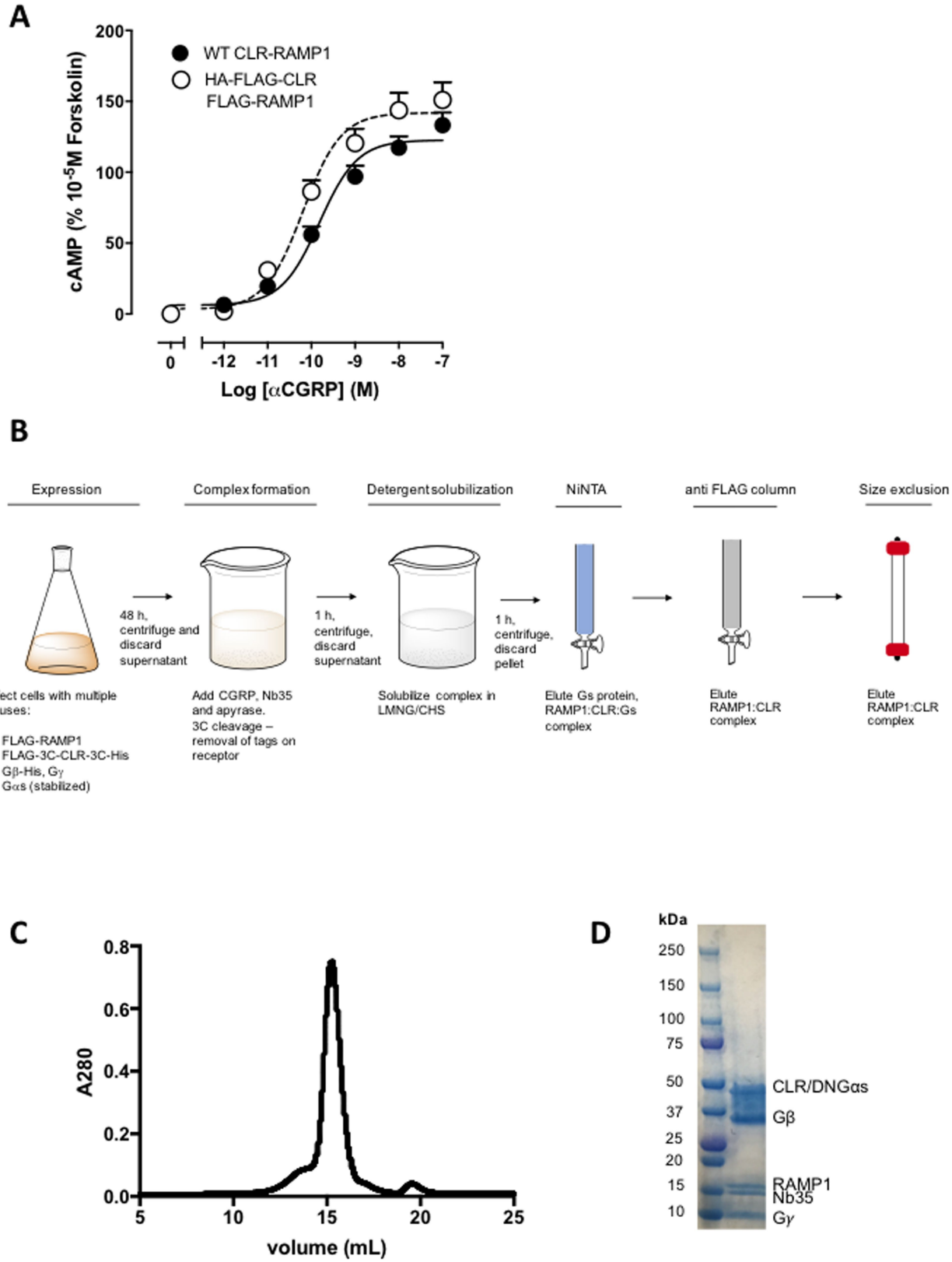
```
  1 MARALCRLPR RGLWLLLAHH LFMTTA
  1 MKTIIALSYI FCLVFADYKD DDDKHGSCQE ANYGALLREL CLTQFQVDME
 51 AVGETLWCDW GRTIRSYREL ADCTWHMAEK LGCFWPNAEV DRFFLAVHGR
101 YFRSCPISGR AVRDPPGSIL YPFIVVPITV TLLVTALVVW QSKRTEGIV
```

CGRP

ACDTATCVTH RLAGLLSRSG GVVKNNFVPT NVGSKAF-NH_2

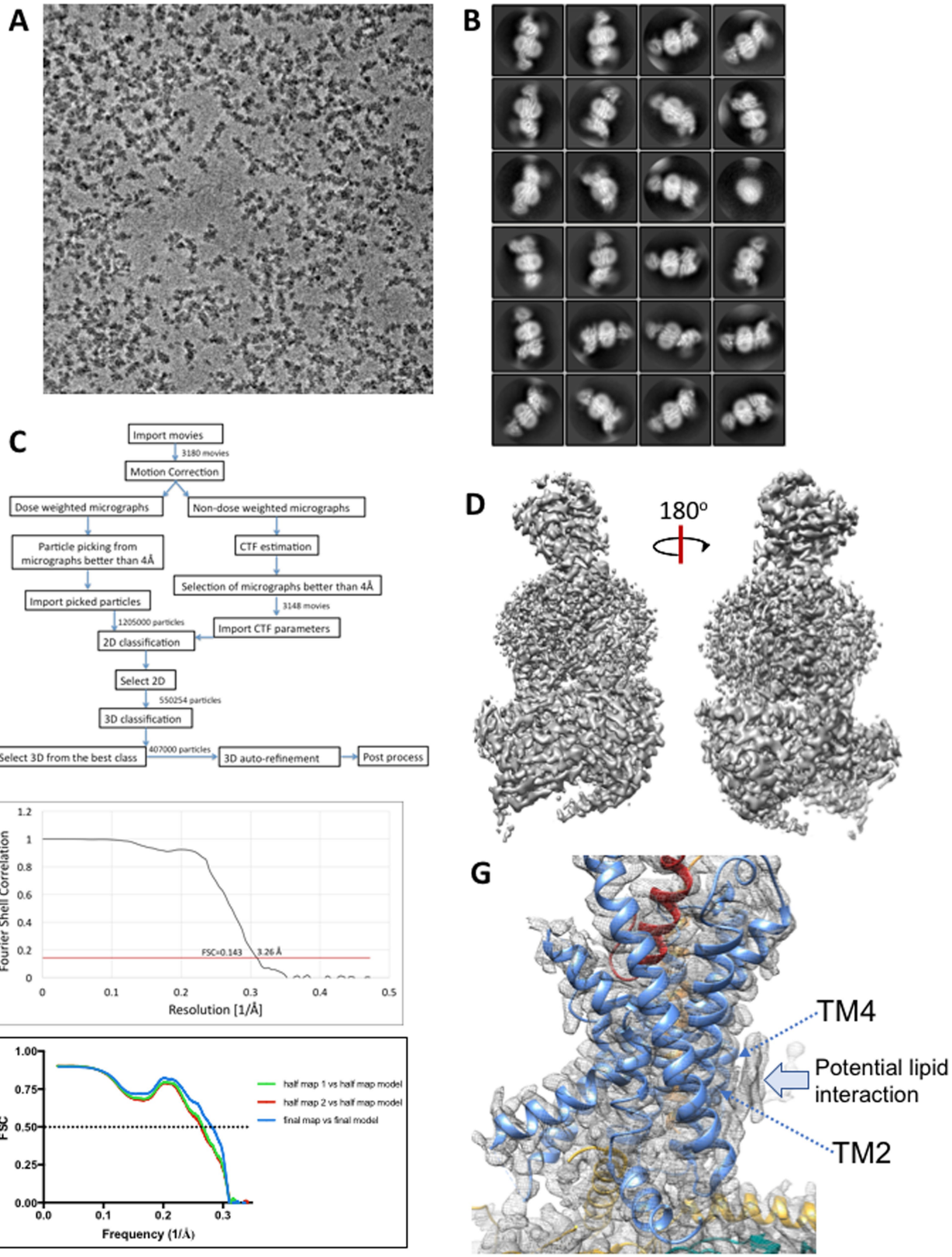
Extended Data Fig. 1 | Amino acid sequences of the CGRP peptide, CLR and RAMP1 constructs used for determination of structure. The sequences are annotated to denote the location of the haemagglutinin (HA) signal sequence (red highlight), 3C cleavage sites (grey highlight), Flag (dark olive-green highlight) and His tags (purple highlight). The substituted sequences of the native proteins are listed above the construct sequences and highlighted in blue. Transmembrane helical domains in CLR and RAMP1 are boxed and highlighted in green. Segments of the proteins that were not resolved in the cryo-EM map are highlighted in yellow. Amino acids for which backbone density was present but there was limited side-chain density were stubbed in the model; these are shown in bold red in the sequences.

© 2018 Springer Nature Limited. All rights reserved.



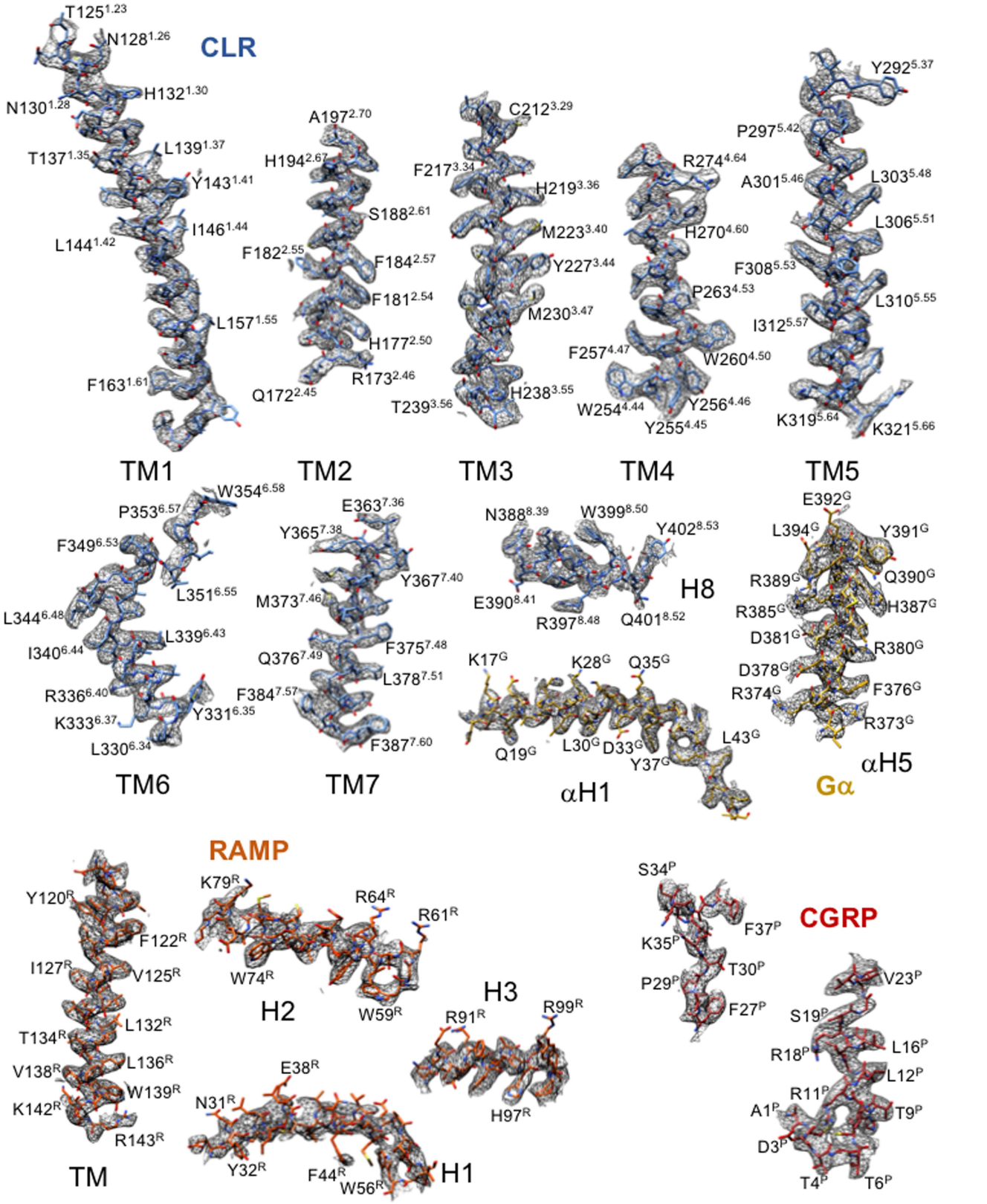
Extended Data Fig. 2 | CGRP receptor pharmacology and purification of the CGRP–CLR–RAMP1–G_s complex. a, Pharmacology of untagged CLR–RAMP1 (wild-type (WT) CLR–RAMP1) and the purification construct (HA–Flag–CLR and Flag–RAMP1), in CGRP-mediated cAMP accumulation assays performed in transiently transfected COS-7 cells. $n = 5$ independent experiments with triplicate repeats; data are mean + s.e.m. **b**, Expression and purification strategy. **c**, Final size-exclusion chromatography elution profile of the complex. **d**, SDS–PAGE and Coomassie blue staining of the size-exclusion chromatography peak, demonstrating the presence of each of the components of the complex.

© 2018 Springer Nature Limited. All rights reserved.



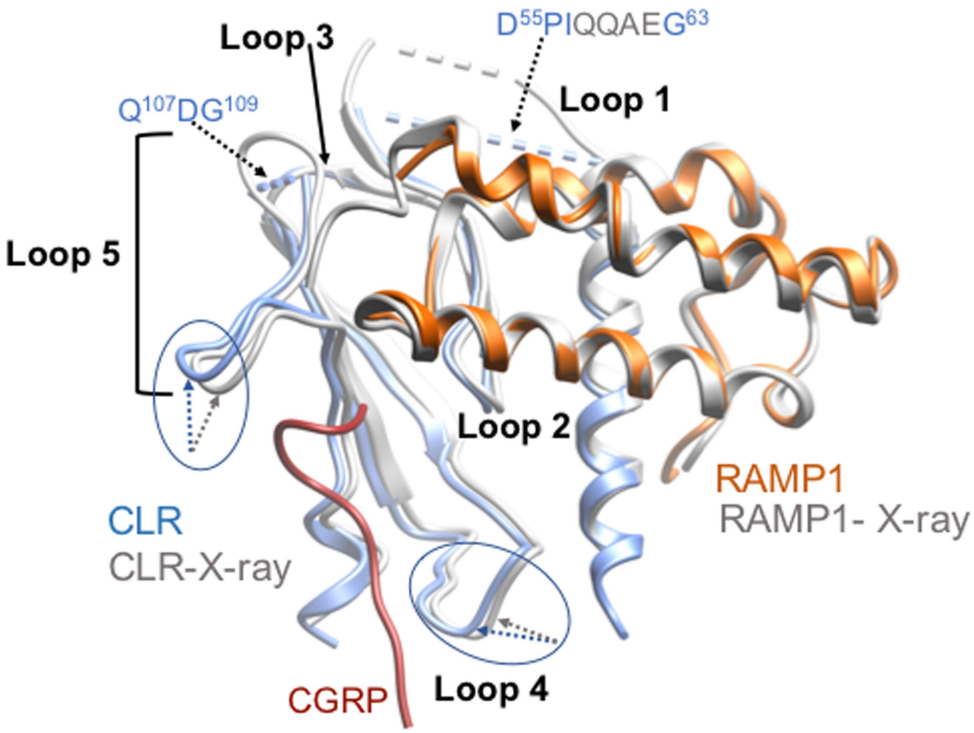
Extended Data Fig. 3 | Volta phase-plate imaging of the CGRP–CLR–RAMP1–G_s complex. a, Volta phase-plate micrograph of the complex (representative of 3,180 movies). High-contrast phase-plate imaging facilitates robust particle selection despite low defocus and tight packing of particles. **b**, RELION 2D class averages. **c**, Workflow for map refinement. **d**, Final 3D cryo-EM map calculated in RELION after auto-refinement and map sharpening. **e**, Gold standard Fourier shell correlation curve; the overall nominal resolution is 3.26 Å. **f**, Model overfitting was evaluated by randomly displacing all atoms by 0.5 Å and refined against one cryo-EM half map. Fourier shell correlation curves were calculated between the resulting model and the half map used for refinement (green); the resulting model and the other half map for cross validation (red), and the final refined model and the full map (blue). **g**, Potential lipid interaction with the base of TM4 and TM2 of CLR.

© 2018 Springer Nature Limited. All rights reserved.



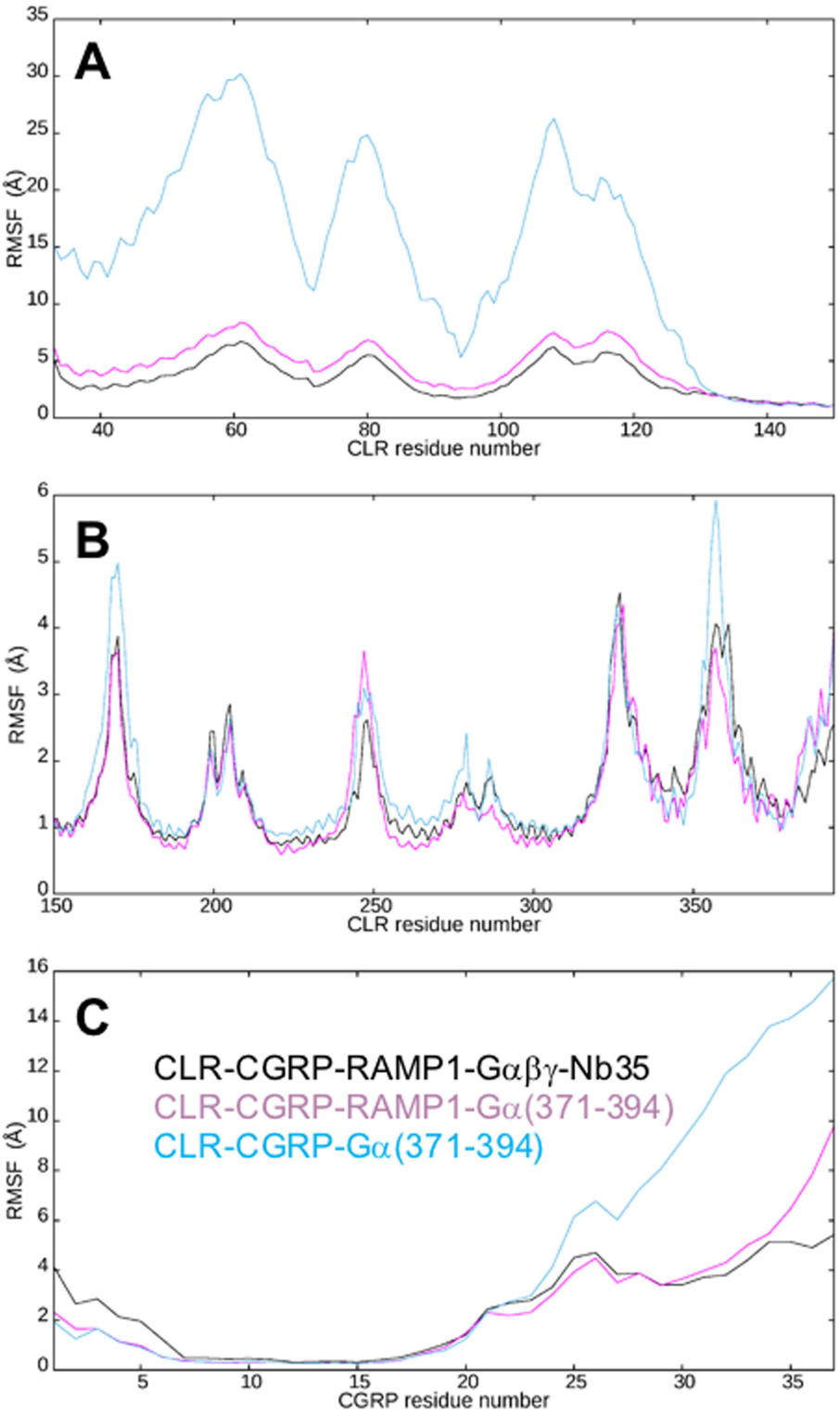
Extended Data Fig. 4 | Atomic-resolution model of the CGRP–CLR–RAMP1–G_s complex in the cryo-EM density map. Cryo-EM density map and model are shown for all seven transmembrane helices and H8 of the receptor, the CGRP peptide (excluding the Lys24-Asn25-Asn26 sequence that was not resolved in the map), the RAMP transmembrane domain and each of the RAMP ECD helices. There was only limited side-chain density for RAMP1 H1, with side chains modelled from rigid-body fitting of the RAMP1 ECD in PDB: 4RWG[12]. The N-terminal (αH1) and C-terminal (αH5) α-helices of the $G\alpha_s$-Ras domain are also shown. Superscript P indicates residues of CGRP.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Alignment of modelled active complex and X-ray structure. Backbone of the ECD of CLR (blue ribbon) and RAMP1 (orange ribbon) from the modelled, active complex, and the structure of the isolated CLR–RAMP1 ECD complex solved by X-ray crystallography[12] (light grey ribbon). The structures were aligned on the RAMP1 ECD. The CLR loops (loops 1–5) are annotated. The CLR loop 1 and loop 5 sequences that were not resolved in the cryo-EM map are indicated by dotted black arrows. Differences in the backbone position of CLR loops 4 and 5 are indicated in blue (active complex) and grey (isolated ECD complex) dotted arrows. The location of the CGRP peptide is shown in dark red.

© 2018 Springer Nature Limited. All rights reserved.

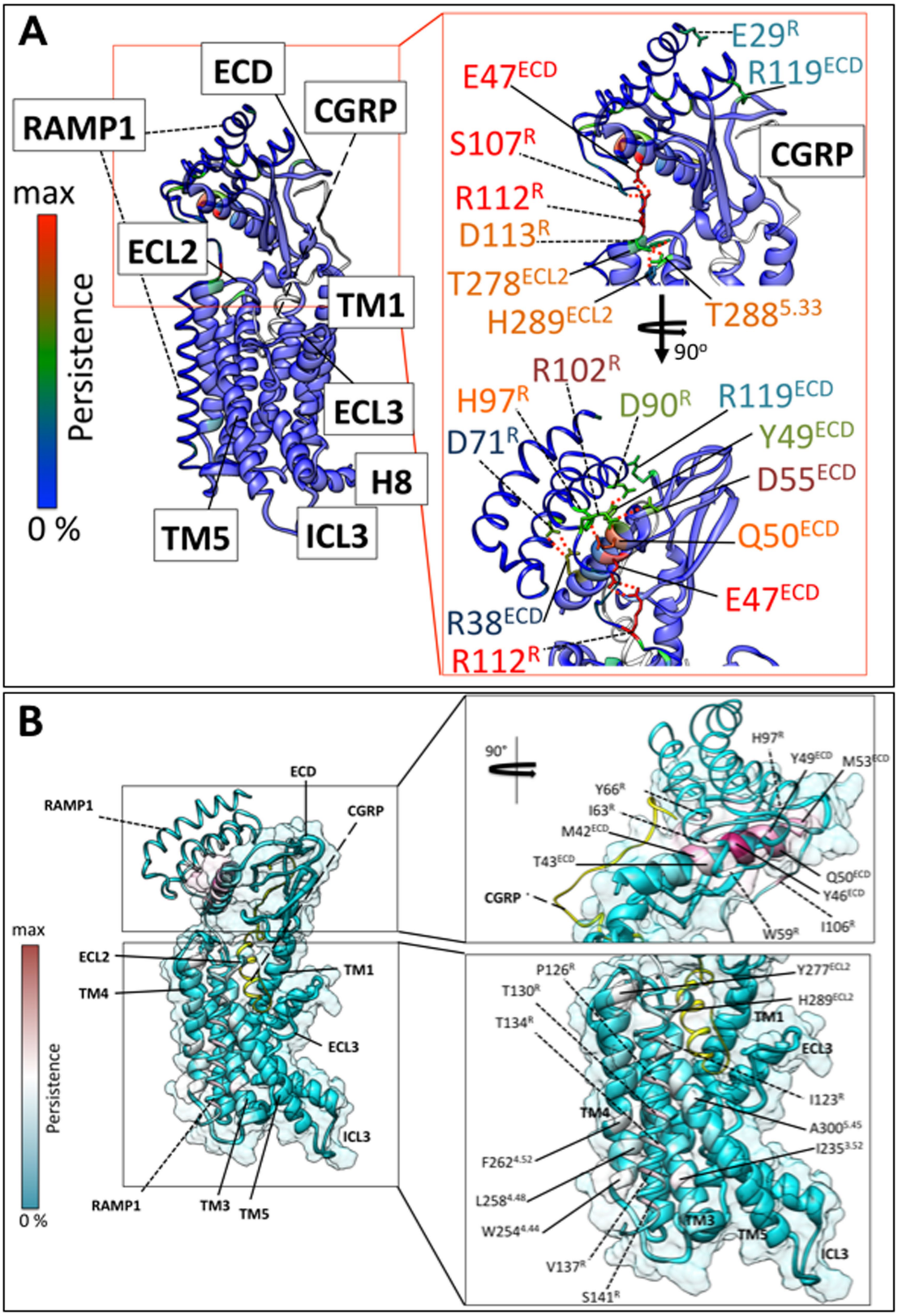


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | RMSF for CGRP and CLR taken from the three simulations. Simulations of CLR–CGRP–RAMP1–Gαβγ–Nb35 (black, 2.4 μs), CLR–CGRP–RAMP1–Gα(371–394) (purple, 2 μs) and CLR–CGRP–Gα(371–394) (blue, 2 μs). **a**, The CLR ECD region. **b**, The CLR transmembrane region. **c**, CGRP (superposed on T6–S17, and therefore valid for the N-terminal half). In general, the missing segments in the cryo-EM density map correspond to regions of high RMSF, and indeed the difficulty of fitting the ECD as a whole is linked to its high RMSF (**a**; Supplementary Videos 2, 3). The segments missing from the ECD ($D55^{ECD}$–$V63^{ECD}$) and ($Q107^{ECD}$–$G109^{ECD}$) correspond to external loop regions furthest removed from the transmembrane domain. Despite their polar nature they displayed no persistent interactions during the molecular dynamics simulations; $D55^{ECD}$–$V63^{ECD}$ displayed the largest backbone RMSF of 8 Å, whereas $Q107^{ECD}$–$G109^{ECD}$ displayed a similarly high RMSF of 7.5 Å. The next-highest RMSF peaks around $A79^{ECD}$–$G81^{ECD}$ and $P115^{ECD}$–$S117^{ECD}$ are just a little lower, but are nonetheless resolved (**a**). Within the transmembrane domain, ICL3 ($H324^{ICL3}$–$S328^{ICL3}$) and ECL3 ($P356^{ECL3}$–$E362^{ECL3}$) both contain missing residues and have a high RMSF above 4.5 Å (**b**). This region displays no persistent interactions during the molecular dynamics simulations, although CGRP does interact with the proximal (non-missing) region of ECL3. The high RMSF values for ICL1 (3.6 Å) and ICL2 (3.6 Å) give rise to stubbed residues ($K167^{2.40}$) and $E248^{ICL2}$–$Q250^{ICL2}$) but the backbone is resolved. For CGRP, the peak in the RMSF around residue 26 (**c**) corresponds to the three highly mobile external residues (Lys24-Asn25-Asn26) in the outward-facing loop that do not interact with CLR (Extended Data Fig. 8); these residues could not be placed from the electron density. These three CGRP residues form a hinge, enabling changes in the orientation of the CLR ECD, especially in the absence of RAMP1; the higher RMSF values C-terminal to this are an artefact of the superposition strategy and the two-domain nature of CLR, but their relative values still hold. The high mobility of some of the extracellular loops is visible in videos (Supplementary Videos 1–3).

© 2018 Springer Nature Limited. All rights reserved.

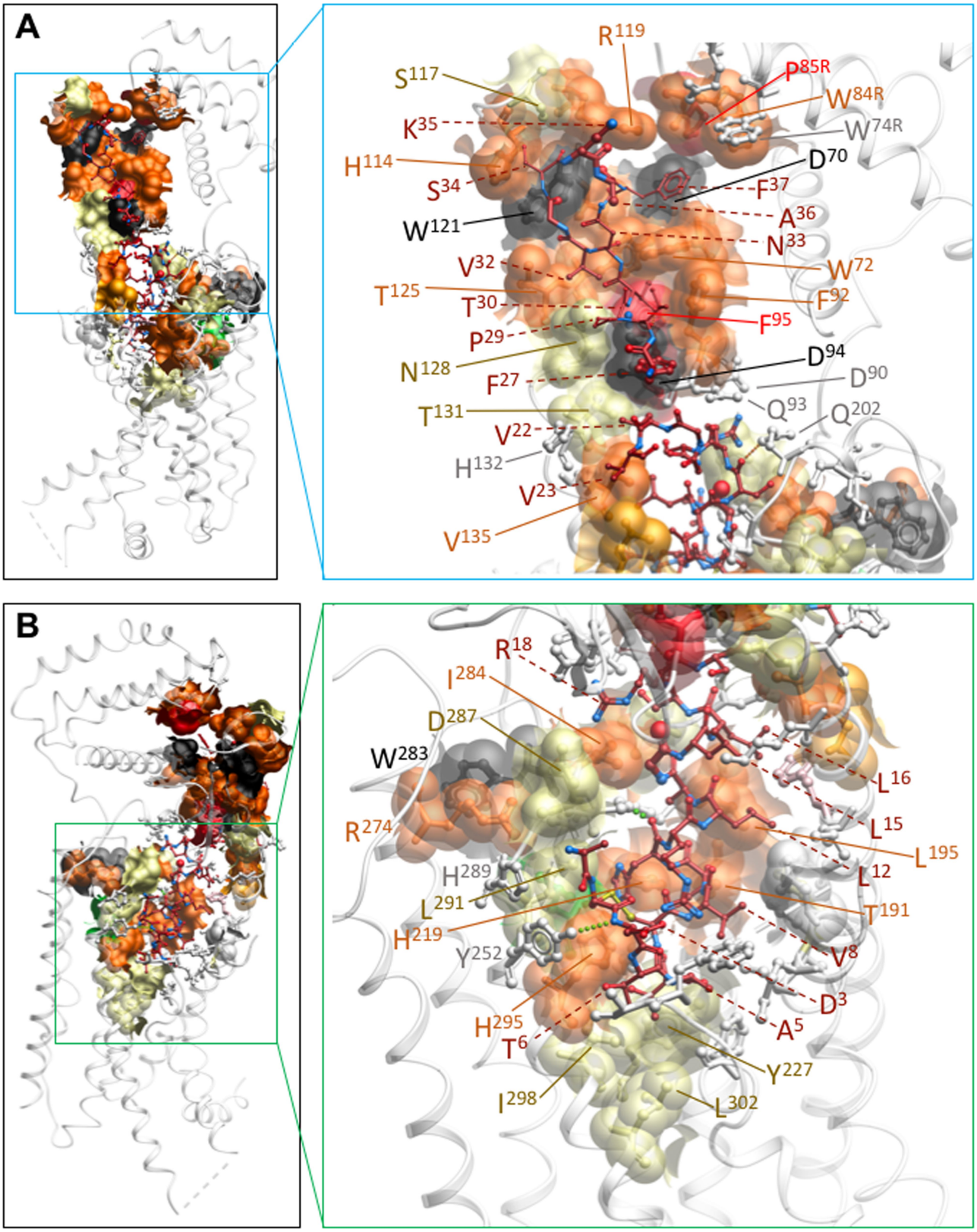


Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

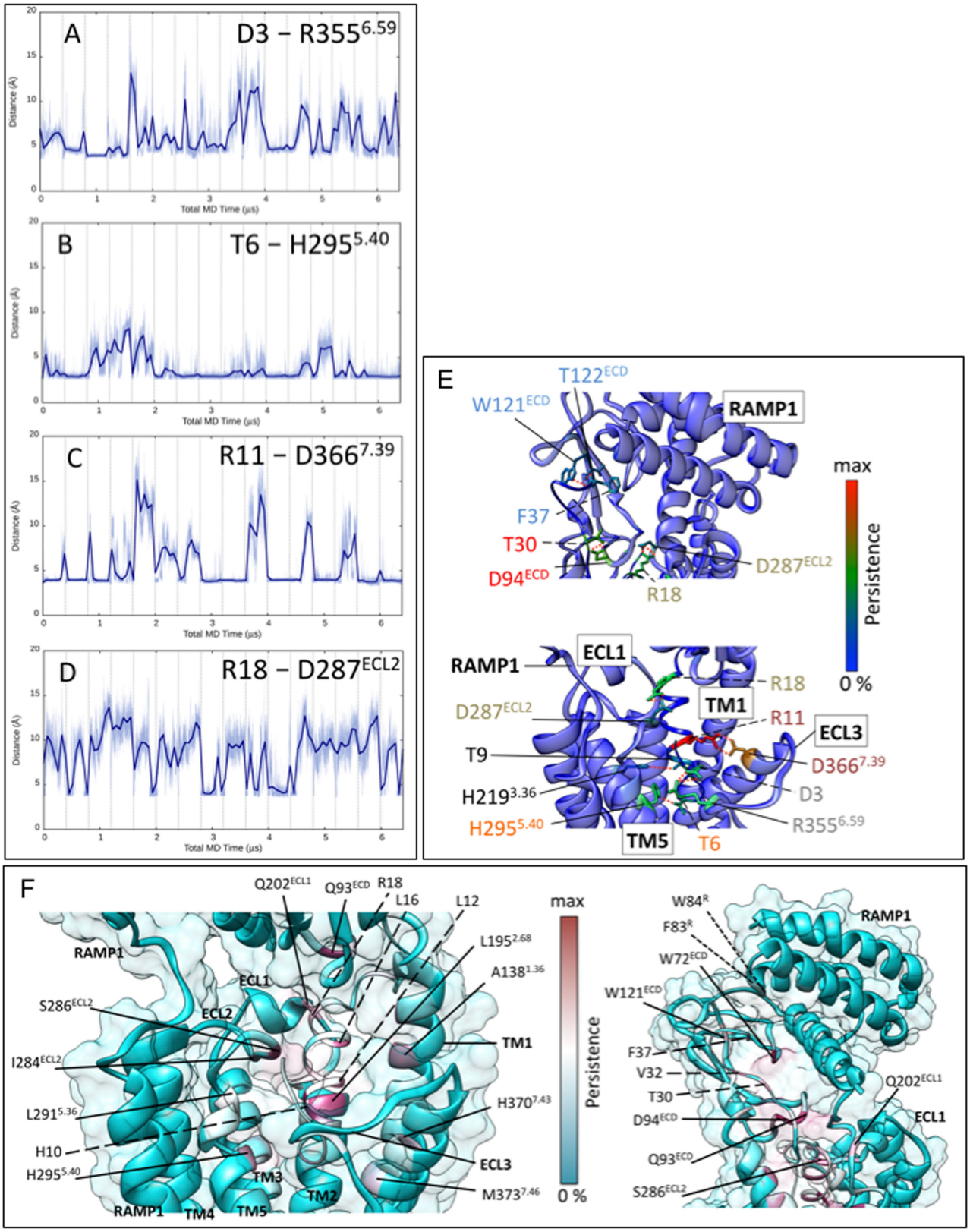
Extended Data Fig. 7 | RAMP1 makes extensive stable interactions with CLR. a, Hydrogen bonds between RAMP1 and CLR during molecular dynamics simulations (6.4 μs). The total persistence is plotted onto the experimental structure according to a rainbow colour scale, with residues that are never involved in dark blue and residues that are highly involved in red. The receptor is shown as a bulky ribbon, RAMP1 as a thin coloured ribbon and the peptide as a thin white ribbon. Key side chains are shown, but for intermittent hydrogen bonds the rotameric state has been modified to show an interaction. Residues forming an interaction network are labelled with the same colour. Left, overall topology of the system. Right top, magnified view of the upper portion of the CLR transmembrane domain and ECD; right bottom, view rotated by 90° on the *z* axis. Hydrogen bonds involved in the RAMP1–CLR interaction, $R112^{R}$–$E47^{ECD}$ and $D113^{R}$–$T288^{ECL2}$/$H289^{ECL2}$ are notable because they link the transmembrane domain to the ECD, and for stabilizing ECL2. Other hydrogen bonds implicated in stabilizing the CLR and RAMP1 ECD interaction include $S107^{R}$–$E47^{ECD}$, $R102^{R}$–$D55^{ECD}$, $H97^{R}$–$Q50^{ECD}$, $D90^{R}$–$Y49^{ECD}$, $D71^{R}$–$R38^{ECD}$ and $E29^{R}$–$R119^{ECD}$. Quantitative data on the persistence of hydrogen bonds during the simulations are reported in Supplementary Table 2. **b**, Contacts between RAMP1 and CLR during simulations (6.4 μs). The total persistence of a residue side chain is plotted onto the experimental structure according to a cyan–maroon colour scale, with residues that are never involved in cyan and residues that are highly involved in maroon. The peptide (italics, dashed line) is depicted as a thin ribbon, whereas the receptor (solid line) is shown as a bulky ribbon and transparent surface. Left, overall topology of the system. Top right, the most-persistent interactions involving RAMP residues and the CLR ECD, $W59^{R}$, $I63^{R}$, $Y66^{R}$, $H97^{R}$ and $I106^{R}$ help to anchor αH3 and the C-terminal RAMP1 regions of αH2 to (residues $M42^{ECD}$, $T43^{ECD}$, $Y46^{ECD}$, $Y49^{ECD}$, $Q50^{ECD}$ and $M53^{ECD}$ of the CLR ECD). Bottom right, the most-persistent hydrophobic interactions between the transmembrane domains of RAMP1 and CLR, namely $I123^{R}$, $P126^{R}$, $T130^{R}$, $T134^{R}$ and $V137^{R}$ (plus $S141^{R}$) help to anchor the RAMP transmembrane helix to CLR (TM3–TM5; CLR residues $Y277^{ECL2}$, $H289^{ECL2}$, $A300^{5.45}$, $I235^{3.52}$, $F262^{4.52}$, $L258^{4.48}$ and $W254^{4.44}$).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Effect of alanine mutagenesis of CLR or RAMP1 on CGRP potency in cAMP accumulation assays. a, ECD alanine mutations. **b**, CLR core alanine mutations. Residues that have been mutated are displayed in x-stick format. Mutated residues with no effect on signalling are coloured off-white. Residues that have significantly altered CGRP signalling[12,23,28,30–32,34,37,38] are also highlighted in transparent CPK representation, coloured according to magnitude of effect. Yellow, <10 fold; dark orange, 10–100 fold; red, 100–1,000 fold; black, >1,000 fold. The backbones of CLR and RAMP (solid lines) are displayed in transparent, off-white coloured ribbon. The CGRP peptide (dashed lines) is represented in x-stick format with carbon atoms in dark red and polar atoms coloured in red or blue.

© 2018 Springer Nature Limited. All rights reserved.

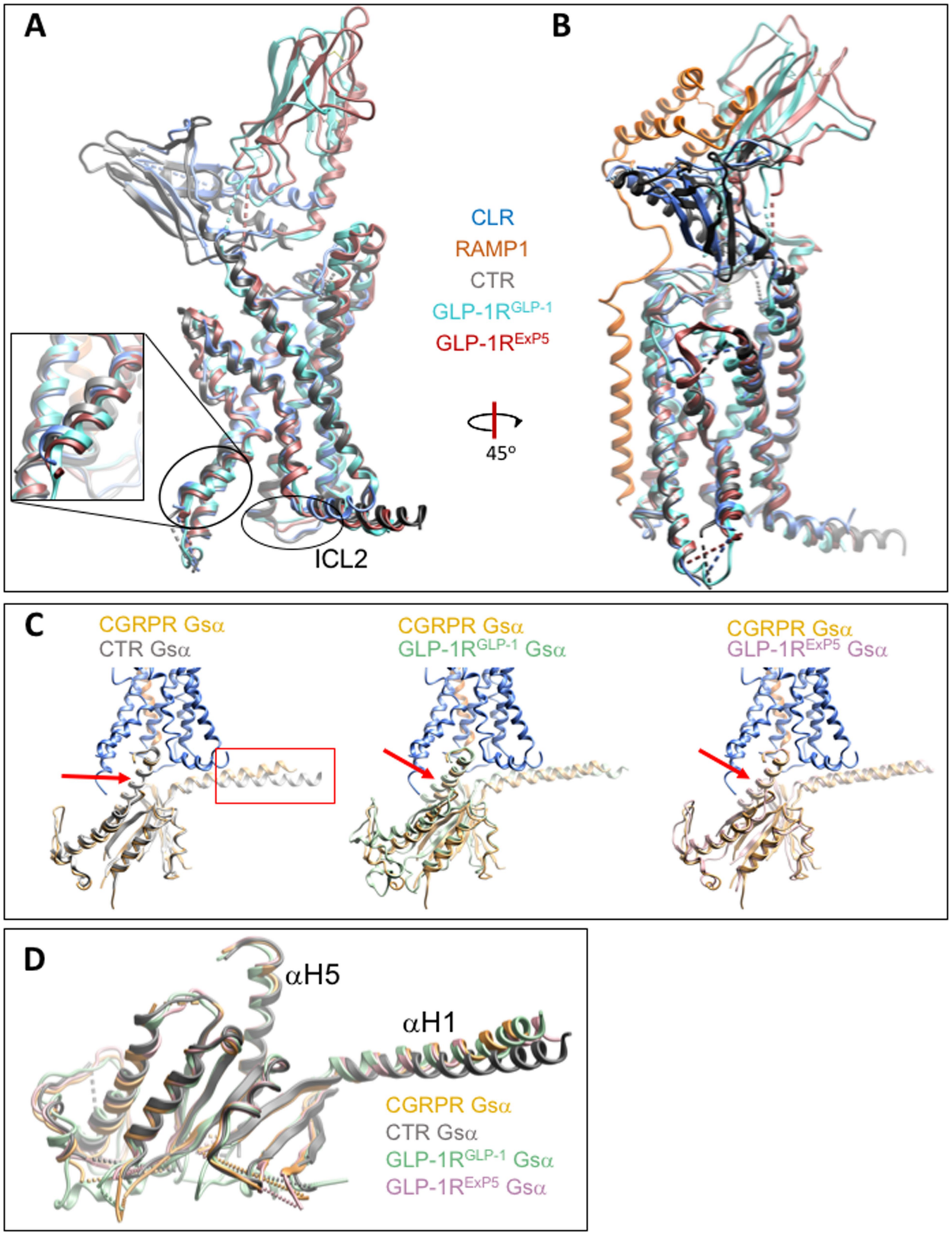


Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | CGRP makes extensive stable interactions with CLR. a–d, Distances between CGRP and CLR residues relevant to key hydrogen bonds. The *x* axis denotes sampling time for the 16 merged molecular dynamics replicas of the whole system (each replica is separated by a vertical dashed line). **a**, Distance between the peptide Asp3 carboxylic carbon and receptor R355$^{6.59}$ guanidinium carbon. **b**, Distance between the peptide Thr6 side-chain oxygen atom and the receptor H295$^{5.40}$ side-chain nitrogen atoms (for each frame, the closest nitrogen to Thr6 was considered). **c**, Distance between the peptide Arg11 guanidinium carbon and the receptor D366$^{7.39}$ carboxylic carbon. **d**, Distance between peptide Arg18 guanidinium carbon and receptor D287^{ECL2} carboxylic carbon. In most cases, the distances corresponding to hydrogen-bond formation are slightly longer than the standard 2.8 Å. **e**, Hydrogen bonds between CGRP and CLR during molecular dynamics simulations (6.4 μs). The total persistence of a residue side chain is plotted onto the experimental structure according to a rainbow colour scale, with residues that are never involved in blue and residues that are highly involved in red. The peptide (italics, dashed line) is depicted as thin ribbon, whereas the receptor (solid line) is shown as bulky ribbon. Key side chains are shown, but for intermittent hydrogen bonds, the rotameric state has been modified to show an interaction. Residues forming an interaction network are labelled with the same colour. Bottom, hydrogen bonds between the CGRP N terminus and the transmembrane bundle of CLR. Top, hydrogen bonds between the CGRP C terminus and the ECD of CLR; quantitative data on the persistence of hydrogen bonds during the simulations are reported in Supplementary Table 3. **f**, Contacts between CGRP and CLR–RAMP1 during molecular dynamics simulations (6.4 μs). The total persistence of a residue side chain is plotted onto the experimental structure according to a cyan–maroon colour scale, with residues that are never involved in cyan and residues that are highly involved in maroon. The peptide (italics, dashed line) is depicted as a thin ribbon, while the receptor (solid line) is shown as a bulky ribbon and transparent surface. Left, contacts between the N terminus of CGRP and the transmembrane bundle of the CLR: highly persistent hydrophobic interactions characterize peptide residues Leu12, Leu16, His10 and receptor residues L195$^{2.68}$, A138$^{1.36}$ and H295$^{5.40}$. Right, contacts between the C terminus of CGRP and the ECD of CLR; highly persistent contacts characterize peptide residues Val32, Thr30, Phe37 and receptor residues Q93^{ECD} and W72^{ECD}. RAMP1 residues F83^{R}, W84^{R} are mainly engaged by CGRP residue Phe37.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Class B GPCRs display similar active state conformations. a, **b**, Alignment of the CGRP–CLR–RAMP1, sCT–CTR, ExP5–GLP-1R and GLP-1–GLP-1R structures (aligned on the transmembrane domains). Regions of divergence between CLR/CTR and GLP-1R are circled. In **a**, RAMP1 has been omitted for clarity. **c**, Position of the $G\alpha_s$-Ras domain in the CTR (left), GLP-1R (GLP-1 bound; middle) and GLP-1R (ExP5 bound; right). The receptor transmembrane domains were aligned. Only the CLR (blue) and RAMP1 (orange) are displayed for clarity. **d**, The $G\alpha_s$-Ras domain from each of the four structures, aligned to the $G\alpha_s$-Ras of the CGRP receptor (CGRPR) complex.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0525-0

Outbursts of luminous blue variable stars from variations in the helium opacity

Yan-Fei Jiang[1]*, Matteo Cantiello[2,3], Lars Bildsten[1,4], Eliot Quataert[5], Omer Blaes[4] & James Stone[3]

Luminous blue variables are massive, evolved stars that exhibit large variations in luminosity and size on timescales from months to years, with high associated rates of mass loss[1–5]. In addition to this on-going variability, these stars exhibit outburst phases, during which their size increases and as a result their effective temperature decreases, typically to about 9,000 kelvin[3,6]. Outbursts are believed to be caused by the radiation force on the cooler, more opaque, outer layers of the star balancing or even exceeding the force of gravity, although the exact mechanisms are unknown and cannot be determined using one-dimensional, spherically symmetric models of stars because such models cannot determine the physical processes that occur in this regime[7]. Here we report three-dimensional simulations of massive, radiation-dominated stars, which show that helium opacity has an important role in triggering outbursts and setting the observed effective temperature during outbursts of about 9,000 kelvin. It probably also triggers the episodic mass loss at rates of 10^{-7} to 10^{-5} solar masses per year. The peak in helium opacity is evident in our three-dimensional simulations only because the density and temperature of the stellar envelope (the outer part of the star near the photosphere) need to be determined self-consistently with convection, which cannot be done in one-dimensional models that assume spherical symmetry. The simulations reproduce observations of long-timescale variability, and predict that convection causes irregular oscillations in the radii of the stars and variations in brightness of 10–30 per cent on a typical timescale of a few days. The amplitudes of these short-timescale variations are predicted to be even larger for cooler stars (in the outburst phase). This short-timescale variability should be observable with high-cadence observations.

We used 60 million CPU hours on the supercomputer Mira, awarded by the Argonne Leadership Computing Facility for the INCITE programme, as well as computational resources from NASA and NERSC (National Energy Research Scientific Computing Center), to solve the three-dimensional radiation hydrodynamic equations[8] and follow the physically realistic evolution of the stellar envelopes of luminous blue variables (LBVs; see Methods for the simulation set-up). These simulations take as inputs the fixed mass of the core of the star and its luminosity from the bottom (inner) boundary of the simulation area, and determine self-consistently the structure of the envelope, its effective temperature and the rate of mass loss. We performed three simulations: the first (which we refer to as T9L6.2) assumes an initial core mass of $M_i = 80M_\odot$, where $M_\odot$ is the mass of the Sun, a luminosity relative to that of the Sun of $\log(L/L_\odot) = 6.2$ and an effective temperature of $T_{eff} = 9{,}000$ K; the second (T19L6.4) assumes the same initial mass, $M_i = 80M_\odot$, $\log(L/L_\odot) = 6.4$ and $T_{eff} = 19{,}000$ K; and the third (T19L6) assumes $M_i = 35M_\odot$, $\log(L/L_\odot) = 6.0$ and $T_{eff} = 19{,}000$ K. Although the three simulations are based on one-dimensional models for stars with different initial masses and at different evolutionary stages (see Methods), they share properties such as density and Eddington ratio (the ratio of radiation to gravitational acceleration) at the peak in opacity due to iron (at around 180,000 K). The main difference between our three simulations is the pressure scale height at the iron opacity peak, which results in different total stellar masses and optical depths above the convective region, and hence different surface temperatures[9].

The locations of the three models in the Hertzsprung–Russell diagram and the stellar-evolution tracks determined from the one-dimensional models are shown in Fig. 1. The histories of spherically averaged radial profiles of density, turbulent velocity, radiation temperature and opacity for run T9L6.2 are shown in Fig. 2. The envelope is convectively unstable at the iron opacity peak[7,9,10], as indicated by the density increasing with radius around that region in the initial hydrostatic structure. Convection takes about 10 dynamical times (about 43 h; where one dynamical time is the time required for the radiation-supported sound waves to travel one pressure scale height in the convective region) to destroy the density inversion (density inversion means that the density increases rather than decreases with radius), which causes high-density clumps to rise, expand and cool to a temperature of less than about 6×10^4 K. Because the density of these cooled clumps is much greater than that of the unclumped gas at the same temperature before the onset of convection, a strong helium opacity peak appears (Fig. 2, bottom). The local radiation acceleration after the helium opacity peak has formed is ten times larger than the gravitational acceleration, which causes a large fraction of the envelope to expand markedly, with most of the gas above that region blown away. The mass flux of unbound gas (positive total energy[11]) leaving our simulation domain can reach an instantaneous rate of around $0.05M_\odot\ \mathrm{yr}^{-1}$. After 400 h, the envelope settles to a steady state(Fig. 2, right). Convection is still operating at around $80.90R_\odot$ (where $R_\odot$ is the radius of the Sun), with a second peak in helium opacity at around $200R_\odot$. Convection also causes envelope oscillations (that is, oscillations in the size or radius of the star) with a typical timescale of a day. The time-averaged location of the photosphere, at which the total Rosseland optical depth to the outer boundary of the simulation box is unity, is at $342.8R_\odot$, as indicated by the dashed blue lines in Fig. 2. At this location, the average radiation temperature is 9.06×10^3 K. The mass-averaged turbulent velocity is only 1% of sound speed deep in the envelope, but becomes supersonic near the photosphere, causing strong shocks and large temperature and density fluctuations near the photosphere. During each oscillation (evident in both the density and the turbulent velocity in the right panels of Fig. 2), part of the mass becomes unbound, with a mass loss rate of about $5 \times 10^{-6}M_\odot\ \mathrm{yr}^{-1}$. This simulation naturally produces a massive star with luminosity, effective temperature and mass loss consistent with an LBV during an outburst. The traditionally used one-dimensional models cannot capture these properties because of two physical processes that can occur only in the three-dimensional simulations: supersonic turbulent motions that provide effective turbulent pressure support to maintain the extended envelope, and radiative and convective energy transport through the turbulent regions around the opacity peaks. In addition, conversion from radiation to kinetic energy is not well captured by traditional mixing-length theory in one-dimensional models[9].

Run T19L6.4 results in a similar evolution history and similar turbulent structures to run T9L6.2, with one snapshot shown in Fig. 3.

[1]Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA, USA. [2]Center for Computational Astrophysics, Flatiron Institute, New York, NY, USA. [3]Department of Astrophysical Sciences, Princeton University, Princeton, NJ, USA. [4]Department of Physics, University of California, Santa Barbara, CA, USA. [5]Astronomy Department and Theoretical Astrophysics Center, University of California at Berkeley, Berkeley, CA, USA. *e-mail: yanfei@kitp.ucsb.edu

© 2018 Springer Nature Limited. All rights reserved.

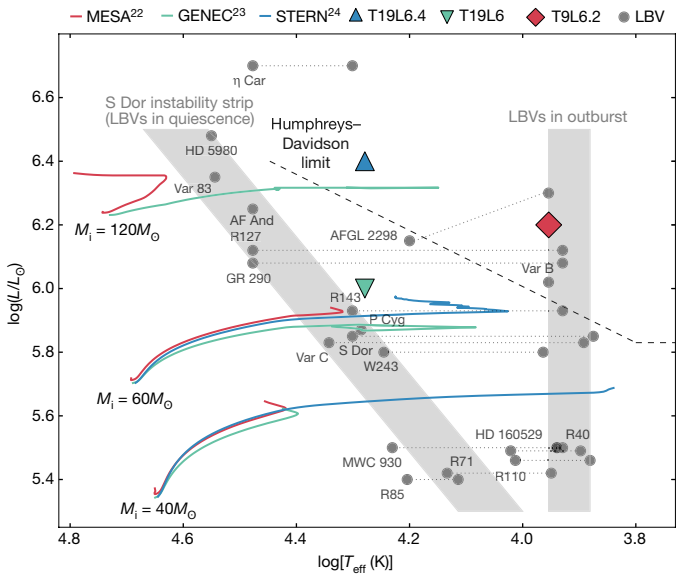


Fig. 1 | Hertzsprung–Russell diagram for LBVs. The shaded areas represent the locations in the diagram where LBVs (grey circles) are most commonly found[5,6]: the diagonal band is the S Dor instability strip (LBVs in quiescence). The vertical band represents LBVs in outburst. Dotted lines indicate observed excursions from quiescence to outburst. The solid red, green and blue lines correspond to main-sequence stellar evolution tracks determined from one-dimensional models with different initial masses as indicated: MESA[21] (modules for experiments in stellar astrophysics), GENEC[22] (Geneva evolution code) and STERN[23], respectively. Our three simulated stars are indicated by coloured polygons. The dashed black line is the Humphreys–Davidson limit[1].

Owing to a smaller pressure scale height and a smaller optical depth across the typical convective element, the gas rising as a result of convection experiences a much smaller temperature change. This results in a much lower value of the opacity at the helium peak compared with run T9L6.2, and thus a smaller total optical depth above the iron-opacity-peak region. Although the luminosity for run T19L6.4 is slightly larger than for run T9L6.2, the less substantial helium opacity peak places the time-averaged location of the photosphere at a smaller radius of $102R_\odot$, with a higher effective temperature of 1.87×10^4 K. This finding confirms that without the helium opacity peak the star will not undergo an outburst and shift into the constant-temperature strip in the Hertzsprung–Russell diagram (see Fig. 1). The presence of a smaller helium opacity peak results in a substantial reduction in the amplitude of the envelope oscillation and a lower associated mass loss rate of around $1 \times 10^{-6} M_\odot$ yr^{-1}.

Finally, run T19L6 has very similar properties to T19L6.4, in particular a comparable value of the pressure scale height at the iron opacity peak. However, this model is calculated for a smaller core mass and a lower luminosity. At the steady state, the envelope has an effective temperature of 1.89×10^4 K, a time-averaged photosphere radius of $63.7R_\odot$ and an episodic mass loss rate associated with envelope oscillations of only around $5 \times 10^{-7} M_\odot$ yr^{-1}. This result confirms that when the iron opacity peak is in a region with a small pressure scale height the effective temperature remains too hot for the helium opacity to become important, and the star stays closer to the S Dor instability strip in the Hertzsprung–Russell diagram (see Fig. 1).

Our simulations predict that LBVs undergoing outbursts should exhibit irregular variability with typical timescales of days. In particular, we expect the variability pattern to be different for massive stars in the S Dor instability strip and during outburst (see Fig. 4). For massive stars with effective temperatures near 9×10^3 K, a substantial helium opacity peak exists in the envelope and causes large-amplitude oscillations. The predicted stellar brightness then varies by a factor of roughly 1.5–2 in a day at the steady state (Fig. 4, top). For stars with hotter effective temperatures of nearly 1.9×10^4 K and a weaker helium opacity peak, the variability at the steady state has a much smaller amplitude. However, the luminosity can still vary by about 20% on timescales of a week to a few weeks, which corresponds to the thermal timescale of the envelope above the iron-opacity-peak region. This kind of variability has been seen in recent high-cadence observations of massive stars[12–14], and the correlation between variability and effective temperature can be tested with future observations. The envelope is loosely bound and dominated by turbulent convection (Fig. 2), so the oscillation at the stellar surface is chaotic. However, there are moments in the evolution of the envelope when the majority of the photosphere is falling back onto the core, as suggested by the integrated luminosity in Fig. 4. This finding

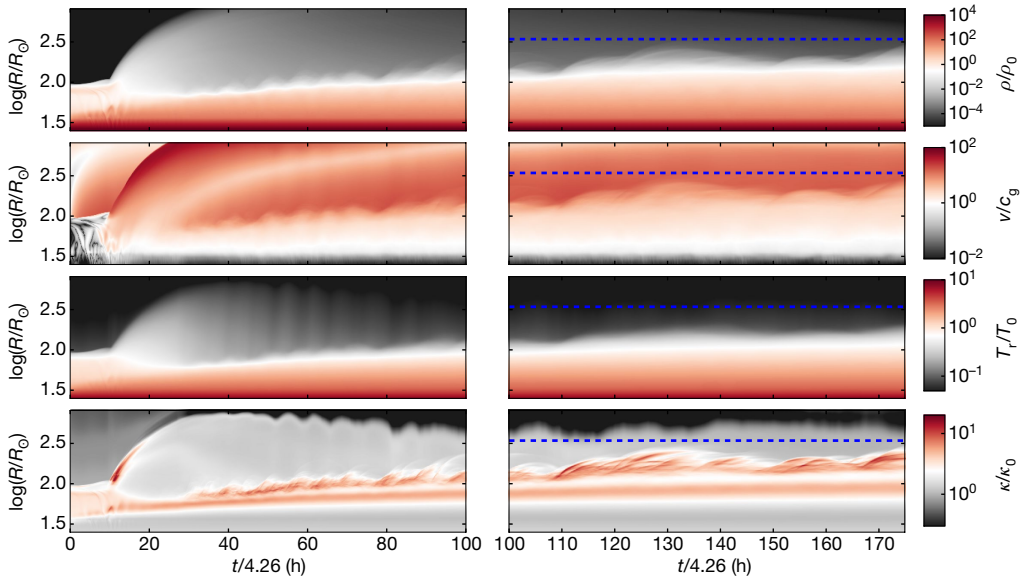


Fig. 2 | Evolution of spherically averaged radial profiles for run T9L6.2. The left and right columns break at $t = 426$ h to separate the initial transition and the steady-state structures. From top to bottom, the colour scales indicate density ρ (scaled by the fudicial density $\rho_0 = 3.6 \times 10^{-9}$ g cm^{-3}), turbulent flow velocity v (scaled by the local isothermal sound speed c_g, which takes the value $c_g = 1.05 \times 10^6$ cm s^{-1} at the photosphere), radiation temperature T_r (scaled by the fudicial temperature $T_0 = 1.67 \times 10^5$ K) and opacity κ (scaled by the fudicial opacity $\kappa_0 = 0.34$ cm^2 g^{-1}). The dashed blue lines indicate where the time-averaged optical depth to the outer boundary of the simulation domain is unity at the steady state. The iron opacity peak is evident in the bottom two panels as the region of larger κ at smaller R, where R is the radial distance from the centre of the star.

© 2018 Springer Nature Limited. All rights reserved.

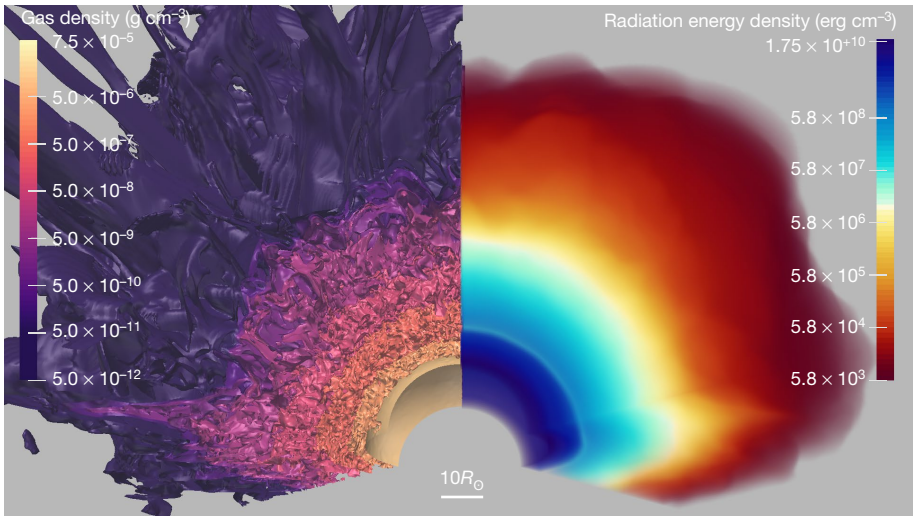


Fig. 3 | A snapshot of the three-dimensional density and radiation energy density from run T19L6.4. The radial range in this plot covers $14.8R_\odot$ (the bottom of the envelope) to $102R_\odot$ (the photosphere). Convection develops at the bottom of the envelope owing to the iron opacity peak at $44.6R_\odot$ (indicated by the region with density of around 5×10^{-8} g cm^{-3}, where small-scale turbulence is observed). The photosphere has large-scale plumes, which cause strong variations in the radiation temperature at the photosphere across the surface of the star. A video of the evolution of this simulation is provided in Supplementary Video 1.

can potentially explain the time-dependent behaviour characterized by P Cygni and inverse P Cygni profiles of some LBVs[13,15].

The mass loss rates that we obtain from our steady-state simulations are broadly consistent with inferred mass loss rates for the quiescent and outburst phases of LBVs[5,15]. We find that the physical mechanism that is responsible for driving the mass loss in LBVs is the interaction of their large radiative flux with opacity peaks that appear in their optically thick envelopes as they expand and cool. Importantly, although the iron opacity peak is strongly metallicity dependent, as long as the turbulent stellar envelope cools to low enough temperatures, the helium and hydrogen opacity peaks will always cause large Eddington factors. This mode of mass loss may therefore be less sensitive to metallicity than are line-driven winds[16]. Traditionally, mass loss due to radiation force on the ultraviolet lines in the optically thin region of the stellar envelope is thought to be the dominant mechanism for winds in these massive stars[17,18], although other models are probably important for outbursts[3,19]. Our work suggests that it is important to study these mechanisms with the turbulent envelope as found by our simulations.

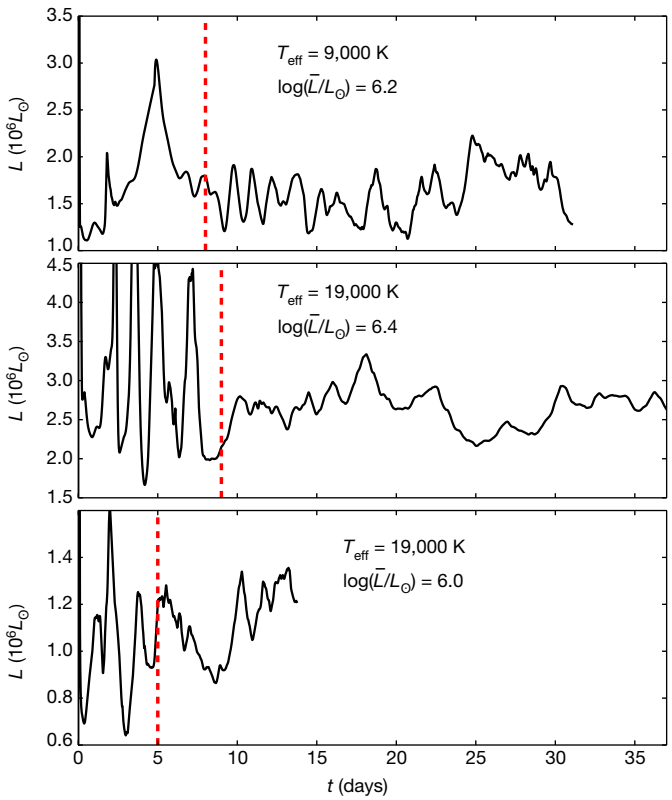


Fig. 4 | Evolution of the total luminosity measured from the outer boundary of the simulation box. Top, T19L6.2 ($M_{\rm i} = 80M_\odot$); middle, T19L6.4 ($M_{\rm i} = 80M_\odot$); bottom, T19L6 ($M_{\rm i} = 35M_\odot$). The vertical dashed red lines indicate the time by which the density inversion in the initial conditions has been removed owing to convection. The effective temperature and average luminosity at the steady state are labelled for each simulation.

The simulations also suggest possible paths for the transition between the S Dor instability strip and the outburst phase in the Hertzsprung–Russell diagram, as indicated by the dotted black lines for the observed LBVs in Fig. 1. When a star expands due to nuclear evolution and a substantial helium opacity peak appears, the star will undergo the first outburst. The amount of mass initially above the iron-opacity-peak region can sustain the associated mass loss rate for only around 10 years; the thermal timescale below the convective region in the envelope is also about 10 years. The star may lose a substantial fraction of the mass above the iron-opacity-peak region via the wind before it has time to adjust to a new structure to keep this large mass loss rate. This will reduce the total optical depth above the iron opacity peak and increase the effective temperature. When the helium opacity peak is reduced substantially, these stars will return to the S Dor instability strip. As the iron opacity peak moves to a deeper region of the envelope on the thermal timescale, this process can repeat. Alternatively, if the massive star is in a binary system, as has been suggested for some LBVs[20], and the companion deposits mass on the surface of the star, then the additional mass will probably be ejected by the massive star, as we found in our initial evolutions for each numerical simulation. This mass ejection could be a trigger of the giant eruptions observed for some LBVs. The detailed properties of this process will need to be studied with future calculations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0525-0.

© 2018 Springer Nature Limited. All rights reserved.

Received: 20 March; Accepted: 17 July 2018;
Published online 26 September 2018.

1. Humphreys, R. M. & Davidson, K. The most luminous stars. *Science* **223**, 243–249 (1984).
2. Lamers, H. J. G. L. M. & Fitzpatrick, E. L. The relationship between the Eddington limit, the observed upper luminosity limit for massive stars, and the luminous blue variables. *Astrophys. J.* **324**, 279–287 (1988).
3. Humphreys, R. M. & Davidson, K. The luminous blue variables: astrophysical geysers. *Publ. Astron. Soc. Pacif.* **106**, 1025–1051 (1994).
4. Owocki, S. P. in *Very Massive Stars in the Local Universe* (ed. Vink, J. S.) 113–156 (Springer, Cham, 2015).
5. Smith, N. Luminous blue variables and the fates of very massive stars. *Phil. Trans. R. Soc. Lond. A* **375**, 20160268 (2017).
6. Smith, N., Vink, J. S. & de Koter, A. The missing luminous blue variables and the bistability jump. *Astrophys. J.* **615**, 475–484 (2004).
7. Paxton, B. et al. Modules for experiments in stellar astrophysics (MESA): planets, oscillations, rotation, and massive stars. *Astrophys. J. Suppl. Ser.* **208**, 4 (2013).
8. Jiang, Y.-F., Stone, J. M. & Davis, S. W. An algorithm for radiation magnetohydrodynamics based on solving the time-dependent transfer equation. *Astrophys. J. Suppl. Ser.* **213**, 7 (2014).
9. Jiang, Y.-F., Cantiello, M., Bildsten, L., Quataert, E. & Blaes, O. Local radiation hydrodynamic simulations of massive star envelopes at the iron opacity peak. *Astrophys. J.* **813**, 74 (2015).
10. Jiang, Y.-F., Cantiello, M., Bildsten, L., Quataert, E. & Blaes, O. The effects of magnetic fields on the dynamics of radiation pressure-dominated massive star envelopes. *Astrophys. J.* **843**, 68 (2017).
11. Jiang, Y.-F., Stone, J. M. & Davis, S. W. A global three-dimensional radiation magneto-hydrodynamic simulation of super-eddington accretion disks. *Astrophys. J.* **796**, 106 (2014).
12. Groh, J. H. et al. On the nature of the prototype luminous blue variable Ag Carinae. I. Fundamental parameters during visual minimum phases and changes in the bolometric luminosity during the S-Dor cycle. *Astrophys. J.* **698**, 1698–1720 (2009).
13. Mehner, A. et al. Spectroscopic and photometric oscillatory envelope variability during the S Doradus outburst of the luminous blue variable R71. *Astron. Astrophys.* **608**, A124 (2017).
14. Conroy, C. et al. A complete census of luminous stellar variability on day to decade timescales. *Astrophys. J.* (in the press); preprint at https://arxiv.org/abs/1804.05860.
15. Wolf, B., Stahl, O. & Inverse, P. Cygni-type profiles in the spectrum of the luminous blue variable S Doradus. *Astron. Astrophys.* **235**, 340–344 (1990).
16. Smith, N. & Owocki, S. P. On the role of continuum-driven eruptions in the evolution of very massive stars and population III stars. *Astrophys. J.* **645**, L45–L48 (2006).
17. Vink, J. S., de Koter, A. & Lamers, H. J. G. L. M. Mass-loss predictions for O and B stars as a function of metallicity. *Astron. Astrophys.* **369**, 574–588 (2001).
18. Smith, N. Mass loss: its effect on the evolution and fate of high-mass stars. *Annu. Rev. Astron. Astrophys.* **52**, 487–528 (2014).
19. Lovekin, C. C. & Guzik, J. A. Pulsations as a driver for LBV variability. *Mon. Not. R. Astron. Soc.* **445**, 1766–1773 (2014).
20. Smith, N. & Tombleson, R. Luminous blue variables are antisocial: their isolation implies that they are kicked mass gainers in binary evolution. *Mon. Not. R. Astron. Soc.* **447**, 598–617 (2015).
21. Choi, J. et al. MESA isochrones and stellar tracks (MIST). I. Solar-scaled models. *Astrophys. J.* **823**, 102 (2016).
22. Brott, I. et al. Rotating massive main-sequence stars. I. Grids of evolutionary models and isochrones. *Astron. Astrophys.* **530**, A115 (2011).
23. Ekström, S. et al. Grids of stellar models with rotation. I. Models from 0.8 to 120 $M_{\odot}$ at solar metallicity (Z=0.014). *Astron. Astrophys.* **537**, A146 (2012).

Acknowledgements We thank J. Insley (ALCF) for helping us to make the image shown in Fig. 3, N. Smith for providing the data for LBVs, and B. Paxton and J. Goodman for conversations and comments. This research was supported in part by the NASA ATP grant ATP-80NSSC18K0560, the National Science Foundation under grant number NSF PHY 11-25915, 17-48958, and in part by a Simons Investigator award from the Simons Foundation (to E.Q.) and the Gordon and Betty Moore Foundation through grant GBMF5076. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) programme. This research used resources of the Argonne Leadership Computing Facility and National Energy Research Scientific Computing Center, which are DOE Offices of Science User Facility supported under contract DE-AC02-06CH11357 and DE-AC02-05CH11231. Resources supporting this work were also provided by the NASA High-End Computing (HEC) programme through the NASA Advanced Supercomputing (NAS) Division at Ames Research Center. The Flatiron Institute is supported by the Simons Foundation

Reviewer information *Nature* thanks C. Lovekin and S. Owocki for their contribution to the peer review of this work.

Author contributions Y.-F.J. ran the simulations, analysed the results and wrote the first draft of the paper. M.C. ran the MESA one-dimensional stellar evolution calculations and made Fig. 1. M.C., L.B., E.Q., O.B. and J.S. read and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0525-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.-F.J.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Typical LBVs have luminosities of $6 \times 10^5 L_\odot$ to about $4 \times 10^6 L_\odot$, and effective temperatures either hotter than roughly 2×10^4 K in quiescence or around 9×10^3 K during outburst (Fig. 1). We use the stellar evolution code MESA[7,24–26] to evolve solar-metallicity stars with initial masses of $M_i = 35 M_\odot$ and $M_i = 80 M_\odot$ to the ranges of luminosity and temperature shown in the Hertzsprung–Russell diagram[9] in Fig. 1. Although these one-dimensional evolution models for stars in this regime are very uncertain, as illustrated by the widely varying end points of the tracks in Fig. 1, they do provide a useful first approximation of the physical conditions in the radiation-dominated envelopes of these stars.

One important feature of this region of luminosity–temperature space is the presence of the iron opacity peak at temperatures of $T \approx 1.8 \times 10^5$ K due to lines of iron-group elements. The opacity of the iron peak is often a factor of a few larger than that from free-electron scattering and can cause the local radiation acceleration to exceed the gravitational acceleration[7]. In this situation, the envelope is unstable to convection at the location of the iron opacity peak[7,27], with the properties of convection and the structure of the envelope depending crucially on how deep within the star the iron opacity peak occurs[9]. The opacity can in principle increase further—to values 100 times that of electron scattering—when the temperature drops below $(1–4) \times 10^4$ K. This arises from helium and hydrogen recombination, but only when the density exceeds about 10^{-9} g cm^{-3}, a value that is much larger than what is realized in one-dimensional hydrostatic structures around this temperature range.

We take the typical luminosity, density and gravity at the location of the iron opacity peak in our one-dimensional models and construct envelopes in hydrostatic and thermal equilibrium in a spherical polar geometry as the initial conditions for our three-dimensional simulations. The whole simulation box initially covers the temperature range 10^4–10^6 K. The bottom boundary of the simulation box has a density and temperature that are orders of magnitude larger than the values at the iron opacity peak and that remain approximately constant. Both OPAL Rosseland and Planck mean opacity tables[28] are included in the simulations to capture the momentum and thermal coupling between the radiation field and the gas[29]. The outer boundary of the simulation box is at least three times the photosphere radius for all of the simulations.

Code availability. The code that we used to do the simulations, Athena++ with the radiative transfer module, is available from the corresponding author on request.

Data availability

The simulation data are available from the corresponding author on request.

24. Paxton, B. et al. Modules for experiments in stellar astrophysics (MESA). *Astrophys. J. Suppl. Ser.* **192**, 3 (2011).
25. Paxton, B. et al. Modules for experiments in stellar astrophysics (MESA): binaries, pulsations, and explosions. *Astrophys. J. Suppl. Ser.* **220**, 15 (2015).
26. Paxton, B. et al. Modules for experiments in stellar astrophysics (MESA): convective boundaries, element diffusion, and massive star explosions. *Astrophys. J. Suppl. Ser.* **234**, 34 (2018).
27. Cantiello, M. et al. Sub-surface convection zones in hot massive stars and their observable consequences. *Astron. Astrophys.* **499**, 279–290 (2009).
28. Iglesias, C. A. & Rogers, F. J. Updated opal opacities. *Astrophys. J.* **464**, 943–953 (1996).
29. Jiang, Y.-F., Stone, J. & Davis, S. W. Super-Eddington accretion disks around supermassive black holes. Preprint at https://arxiv.org/abs/1709.02845 (2017).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0478-3

A topological source of quantum light

Sunil Mittal[1,2]*, Elizabeth A. Goldschmidt[1,3] & Mohammad Hafezi[1,2,4]

Quantum light is characterized by distinctive statistical distributions that are possible only because of quantum mechanical effects. For example, single photons and correlated photon pairs exhibit photon number distributions with variance lower than classically allowed limits. This enables high-fidelity transmission of quantum information and sensing with lower noise than possible with classical light sources[1,2]. Most quantum light sources rely on spontaneous parametric processes such as down-conversion and four-wave mixing[2]. These processes are mediated by vacuum fluctuations of the electromagnetic field. Therefore, by manipulating the electromagnetic mode structure, for example with dispersion-engineered nanophotonic systems, the spectrum of generated photons can be controlled[3–7]. However, disorder, which is ubiquitous in nanophotonic fabrication, causes device-to-device spectral variations[8–11]. Here we realize topologically robust electromagnetic modes and use their vacuum fluctuations to create a quantum light source in which the spectrum of generated photons is much less affected by fabrication-induced disorder. Specifically, we use the topological edge states realized in a two-dimensional array of ring resonators to generate correlated photon pairs by spontaneous four-wave mixing and show that they outperform their topologically trivial one-dimensional counterparts in terms of spectral robustness. We demonstrate the non-classical nature of the generated light and the realization of a robust source of heralded single photons by measuring the conditional antibunching of photons, that is, the reduced likelihood of photons arriving together compared to thermal or laser light. Such topological effects, which are unique to bosonic systems, could pave the way for the development of robust quantum photonic devices.

Spontaneous four-wave mixing (SFWM) is a third-order nonlinear process in which two pump photons at frequency ω_p are annihilated, and two daughter photons, called signal and idler, are generated at frequencies ω_s and ω_i. The spectra of the generated signal and idler photons as well as their correlations are dictated by energy and momentum conservation relations, $2\omega_p = \omega_s + \omega_i$ and $2k_p = k_s + k_i$, where k indicates the momenta of the respective fields. The spectrum is further constrained by the electromagnetic mode structure: that is, the density of states or, equivalently, the dispersion relation $\omega(k)$, which governs the propagation of the pump, signal and idler photons. Recently, the use of nanophotonic systems such as toroidal and ring resonators has provided a compact and scalable route to manipulate the electromagnetic mode structure and hence to implement spectrally engineered sources of correlated photons[2–4]. For example, coupled ring resonator arrays can be used to control the number of spectral modes[7] as well as to enhance the rate of photon pair generation, without compromising their bandwidth[6,12]. However, nanophotonic systems are invariably disposed to fabrication disorder, which can alter the dispersion of the photonic modes in an unpredictable fashion[8–10] and can result in randomness in the spectrum of photons generated by different devices. This randomness limits the scalability of such sources for practical applications in quantum communication and information processing, which often require multiple sources with identical spectra, for example in multi-photon interference[11].

At the same time, the introduction of topological protection in photonic systems has led to the development of a new class of devices that are inherently robust against disorder[13–19]. This robustness can be attributed to the presence of unidirectional, back-reflection-free edge states in these systems. Edge states are characterized by topologically invariant integers[20–22], and photonic transport through these states is therefore protected against local disorder[14,19,23]. Edge states have been used to demonstrate, for example, robust optical delay lines[16,23], reconfigurable photonic pathways[19] and topological lasers[24,25]. However, such demonstrations have so far been confined to the classical regime.

In this work, we use topology for spectral engineering of the quantum fluctuations of the electromagnetic vacuum and implement a robust source of correlated photon pairs generated by SFWM. In particular, we exploit the linear dispersion associated with edge states for efficient phase-matching and show that the photon pair generation is enhanced when the pump, as well as the signal and idler fields, corresponds to edge modes of the system. We demonstrate correlations between the signal and idler photons beyond what is possible with classical sources and show conditional antibunching of photons, confirming the quantum nature of our source and its operation as a source of heralded single photons. More importantly, using measurements over many devices, we show that the robustness of such topological spectral engineering manifests as a robustness in the spectrum of generated photons and that our topological source outperforms a similarly designed topologically trivial source of correlated photons. From a fundamental perspective, our scheme is similar to recent theoretical proposals[26,27] that investigated second- and third-order nonlinearity in topological edge states, respectively. These particle-non-conserving topological photonic systems have no fermionic counterparts.

Our system consists of a two-dimensional (2D) square lattice of ring resonators, positioned at the lattice sites, for which the non-interacting part of the photon dynamics is governed by the integer quantum Hall model (Fig. 1a)[15,16]. A uniform synthetic magnetic field is synthesized by using link rings to couple the neighbouring site rings such that a photon hopping from one lattice site to its neighbour experiences a position- and direction-dependent hopping phase. The tight-binding Hamiltonian describing the linear evolution of the pump, signal and idler photons in the system is given as

$$H_L = \sum_{m,n} \omega_{0\mu} a^\dagger_{m,\mu} a_{m,\mu} - J_{m,n}(a^\dagger_{m,\mu} a_{n,\mu} e^{-i\phi_{m,n}} + a^\dagger_{n,\mu} a_{m,\mu} e^{+i\phi_{m,n}}) \quad (1)$$

Here $\mu \in \{p, s, i\}$ refers to the pump, signal or idler fields, and $a^\dagger_{m,\mu}$ is the corresponding photon creation operator at a lattice site $m = (m_x, m_y)$, with frequency ω_μ. $J_{m,n}$ is the hopping rate of photons between lattice sites m, n and is non-zero only for the nearest-neighbour sites. $\phi_{m,n} = \phi m_y \delta_{m_x, n_x+1} \delta_{m_y, n_y}$ is the hopping phase between lattice sites and results in a uniform synthetic magnetic field flux ϕ per plaquette (highlighted by the dashed white line in Fig. 1a). The energy spectrum of this Hamiltonian can be probed using transmission spectroscopy. For the chosen magnetic field flux $\phi = \pi/2$, the transmission spectrum consists of two edge bands at $\omega_\mu - \omega_{0\mu} \approx \pm 1.5J$, separated by a bulk band centred at $\omega_\mu - \omega_{0\mu} \approx 0$ (Fig. 1c). The edge bands are occupied by the topological edge states, which are confined to the lattice boundary and circulate around the lattice in clockwise and anticlockwise directions,

[1]Joint Quantum Institute, NIST/University of Maryland, College Park, MD, USA. [2]IREAP/Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. [3]US Army Research Laboratory, Adelphi, MD, USA. [4]Department of Physics, University of Maryland, College Park, MD, USA. *e-mail: mittals@umd.edu

© 2018 Springer Nature Limited. All rights reserved.

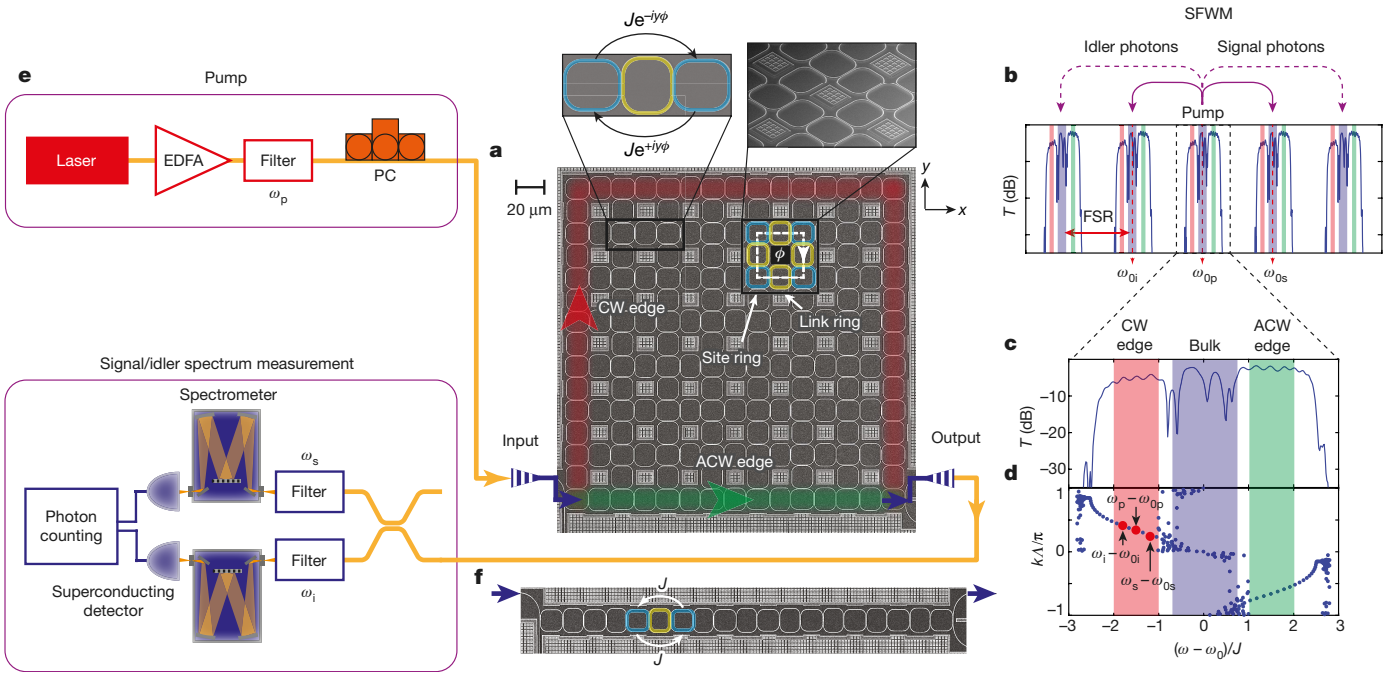


Fig. 1 | Schematic of the experimental set-up. a, Scanning electron microscope (SEM) image of an 8 × 8 lattice of site-ring resonators (cyan), coupled using link rings (yellow). Photons acquire a non-zero phase ϕ when they circulate around a plaquette of four site rings and four link rings (highlighted by the dashed white line). Insets show two site rings coupled by a link ring, and a plaquette. The paths followed by clockwise (CW) and anticlockwise (ACW) edge modes are highlighted in red and green, respectively. **b**, The transmission spectrum of the device repeats after one FSR. Correlated signal and idler photons are generated in longitudinal modes (of individual resonators) located symmetrically around the pump mode (centred at ω_{0p}). We choose the two modes one FSR above and below the pump mode, centred at frequencies ω_{0s} and ω_{0i}, for collection of signal and idler photons. **c**, Simulated transmission (T) spectrum of an 8 × 8 lattice, in a given band. Two edge bands (shaded red and green) are separated by a bulk band (shaded blue). **d**, Simulated dispersion curve showing linear dispersion for the edge modes. Here k is the photon momentum and Λ is the lattice constant such that $k\Lambda$ is the phase between two neighbouring site rings on the edge. Efficient phase-matching occurs when the pump as well as signal and idler frequencies correspond to edge modes. **e**, Schematic of the pump and the spectral measurement set-up. EDFA, erbium-doped fibre amplifier; PC, polarization controller. **f**, SEM image of a topologically trivial 1D array of ten site-ring resonators (cyan), coupled using link rings (yellow).

respectively (Fig. 1a)[16]. Furthermore, edge states are well described by a linear dispersion relation (Fig. 1d)[15,22]. In contrast, states in the bulk band occupy the bulk of the system and do not have a well-defined momentum. Note that this edge/bulk band structure repeats after every free spectral range (FSR)—that is, the frequency spacing between consecutive longitudinal modes of the individual ring resonators (Fig. 1b). Consequently, the pump, signal and idler fields can occupy different longitudinal modes with resonance frequencies denoted by $\omega_{0\mu}$.

To generate correlated photon pairs in this system, we use the third-order nonlinearity of silicon and realize SFWM. This nonlinear four-photon interaction is described by the Hamiltonian

$$H_{NL} = \eta \sum_m (a^\dagger_{m,s} a^\dagger_{m,i} a_{m,p} a_{m,p} - a^\dagger_{m,p} a^\dagger_{m,p} a_{m,s} a_{m,i}) \quad (2)$$

where η is the strength of the SFWM and depends on the material and ring waveguide properties[28,29]. The signal and idler modes are initially in the vacuum state when the input pump photons enter the system. However, the nonlinear interaction coherently adds or removes photon pairs from these vacuum modes and leads to generation of non-classical fields with intensity and spectral correlations between signal and idler photons[26]. Furthermore, because of energy conservation, correlated signal and idler photon pairs are generated in longitudinal modes (of individual resonators) located symmetrically on either side of the pump mode[28,29]. We choose signal and idler modes a single FSR above and below the pump mode, with resonance frequencies denoted by ω_{0s}, ω_{0i} and ω_{0p}, respectively (Fig. 1b). This choice allows us to effectively filter out the pump photons at the device output and also minimize the phase-mismatch arising from the waveguide and material dispersion[3,4].

In our experiment, we pump the lattice in one of the longitudinal modes using a tunable continuous-wave laser and measure the spectrum of generated photons. Figure 2a plots the linear pump transmission spectrum (see Extended Data Fig. 3 for more details) and Fig. 2b plots $I(\omega_s, \omega_p)$, the intensity of generated signal photons at frequency ω_s as we tune the pump frequency ω_p. For a continuous-wave pump, measurement of $I(\omega_s, \omega_p)$ is equivalent to a measurement of the joint-spectral intensity which is commonly used to characterize the spectral correlations between generated photons (see Extended Data Fig. 4 and refs [11,29]). First, we observe that the maximum number of photons is generated when the lattice is pumped in the clockwise edge band, at $\omega_p - \omega_{0p} \approx -1.5J$. Second, with a clockwise edge band pump, the spectrum of generated signal photons is predominantly confined to the clockwise edge band. This limited spectral distribution of signal photons can be seen more clearly with a normalized spectrum, the horizontal cross-section of $I(\omega_s, \omega_p)$, at $\omega_p - \omega_{0p} \approx -1.5J$ (Fig. 2i). Furthermore, as a consequence of energy conservation, idler photons also exhibit a similar narrow spectrum centred at the clockwise edge band: that is, $\omega_i - \omega_{0i} = 2(\omega_p - \omega_{0p}) - (\omega_s - \omega_{0s}) \approx -1.5J$ (Fig. 2j). This enhanced and spectrally limited generation of correlated photon pairs in the edge band is a result of the linear dispersion of edge modes which naturally satisfies the phase-matching criteria, and of good spatial overlap between the pump, signal and idler photons when they are confined to the lattice boundary. Our simulation results for the generated photon spectra agree well with our experimental observation (Fig. 2k). We observe a similar spectrally limited generation of correlated photons when the pump frequency is in the anticlockwise edge band (Fig. 2c–e). However, the propagation distance from the input to the output port is much shorter for the anticlockwise edge modes than for the clockwise edge modes, and therefore the intensity of generated photons is much weaker (Fig. 1a).

© 2018 Springer Nature Limited. All rights reserved.

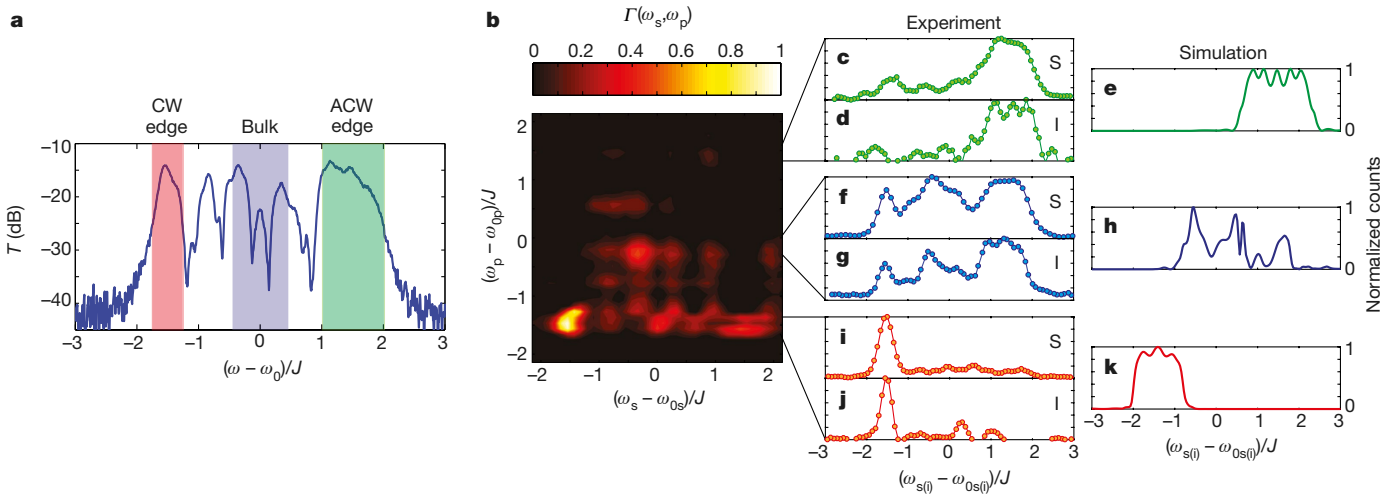


Fig. 2 | Spectral distribution of the generated photons. a, Measured transmission spectrum for an 8 × 8 lattice. The edge and bulk bands are highlighted in colour. **b**, $\Gamma(\omega_s, \omega_p)$, the intensity (normalized to unity) of generated signal photons at frequency ω_s as a function of pump frequency ω_p. The maximum number of photons are generated when the pump as well as the signal and idler frequencies are in the clockwise edge band ($\omega - \omega_0 \approx -1.5J$). **c**, Spectrum (normalized to unity) of signal (S) photons, that is, horizontal cross-section of $\Gamma(\omega_s, \omega_p)$, at $\omega_p \approx 1.5J$. With the pump in the anticlockwise edge band, the spectrum of generated signal photons is also limited to the anticlockwise edge band. **d**, Spectrum of idler (I) photons, with the pump in the anticlockwise edge band. Because of energy conservation, idler photons are also generated predominantly in the anticlockwise edge band. **e**, Simulation results for the spectrum of generated photons match well with experimental observations. **f**–**h**, Corresponding results for the pump in the bulk band. The signal and idler photons are generated throughout the spectrum of the lattice. Also, the simulation results do not match the observation because of the fabrication disorder in the experimental system. **i**–**k**, Signal and idler spectra when the system is pumped along the clockwise edge band, again showing spectrally confined generation of photons in the edge band.

In contrast to edge modes, bulk modes do not have a well-behaved dispersion (see Fig. 1d), and their intensity distribution in the lattice changes even for very small changes in the excitation frequency[16]. Therefore, in the bulk band, there is a phase mismatch between the pump, signal and idler photons, and their spatial overlap is also limited. As a result, the SFWM efficiency is low, and photon pairs are generated throughout the transmission band of the lattice (Fig. 2f–h). Moreover, the experimental and simulation results for the bulk band pump do not match. This is because our experimental system has fabrication disorder, and the bulk band is not robust against disorder. In contrast, good agreement between the observed and simulated results for the edge states indicates their robustness against disorder.

To characterize the non-trivial nature of correlations between generated photons, we measure the second-order cross-correlation function, $g_{s,i}^{(2)}(\tau)$, which is the normalized probability of detecting signal and idler photons separated by time τ (see Methods and refs [2,5]). For two uncorrelated sources, $g^{(2)} = 1$ for all τ. In contrast, we observe a maximum $g_{s,i}^{(2)} \approx 80$ at $\tau = 0$ (Fig. 3a). We integrate $g_{s,i}^{(2)}(\tau)$ over the peak at $\tau = 0$ to obtain the ratio of coincidence to accidental counts (CAR), which is analogous to the signal-to-noise ratio of a source. Our source achieves a CAR ≈ 42 (Fig. 3c), higher than other similar sources using single resonators[4,5] and coupled resonators[6], for which CAR values of approximately 30 and 10 were reported, respectively. This clearly indicates that the signal and idler photons are strongly correlated: that is, the detection of a signal photon heralds the arrival of an idler photon and vice versa. Furthermore, we verified that the coincidence count rate between signal and idler photons increases as the square of the pump power (Fig. 3b) and CAR drops inversely with the coincidence rate (Fig. 3c), as expected for SFWM interaction[3–5].

Next, using a Hanbury Brown–Twiss set-up, we measure the conditional (heralded) autocorrelation function, $g_{a,h}^{(2)}(\tau)$, for signal photons, conditioned on the detection of idler photons (see Methods and refs [2,5]). Classical light sources are characterized by $g_a^{(2)}(0) \geq 1$ where the inequality holds for sources with bunched photons (such as thermal light), and $g_a^{(2)}(0) = 1$ when there are no correlations between arrival times of photons (as in lasers). Quantum light sources, such as single photons, are distinguished by $g_a^{(2)}(0) < 1$, which means that the photons are anti-bunched. We observe a conditional $g_{a,h}^{(2)}(0) = 0.20(8)$, which shows antibunching and confirms that we have realized a topological source of heralded single photons (Fig. 3d).

Edge states are topologically protected, quasi-one-dimensional (1D) waveguides confined to the lattice boundary. Therefore, to benchmark the robustness of these edge channels, we compare them with the topologically trivial 1D coupled resonator optical waveguides (CROWs; Fig. 1f)[23,30]. The main advantage of CROWs over single-ring devices is that they increase the length of SFWM interaction and therefore the intensity of generated photons, without reducing their bandwidth[6,7,12]. However, unlike edge states, CROWs are not protected against disorder, which can affect the photonic mode structure (see ref. [9] and Extended

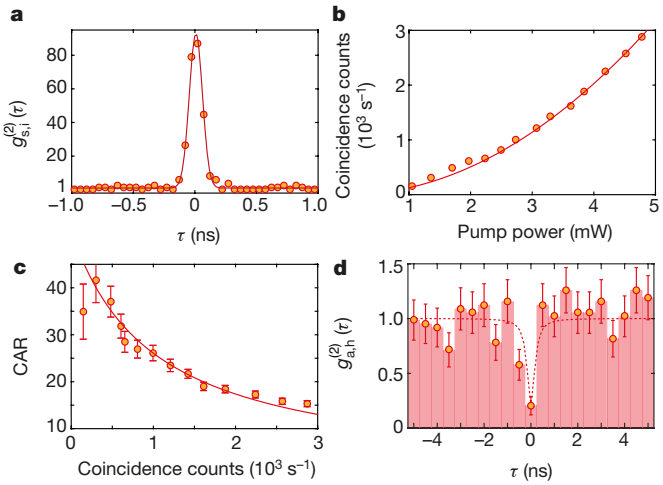


Fig. 3 | Source characterization. a, Histogram for the cross-correlation function $g_{s,i}^{(2)}(\tau)$ between signal and idler photons, with a pump power of approximately 1.4 mW. **b**, Coincidence count rate at the device output (adjusted for coupling losses), as a function of pump power. **c**, CAR as a function of the coincidence count rate. **d**, Histogram for the conditional (heralded) autocorrelation function $g_{a,h}^{(2)}$, with $g_{a,h}^{(2)}(0) = 0.20(8)$. The error bars in **b** and **c** are calculated assuming Poissonian statistics for photon counts. Solid lines are fits to the data and the dashed line is to guide the eye.

© 2018 Springer Nature Limited. All rights reserved.

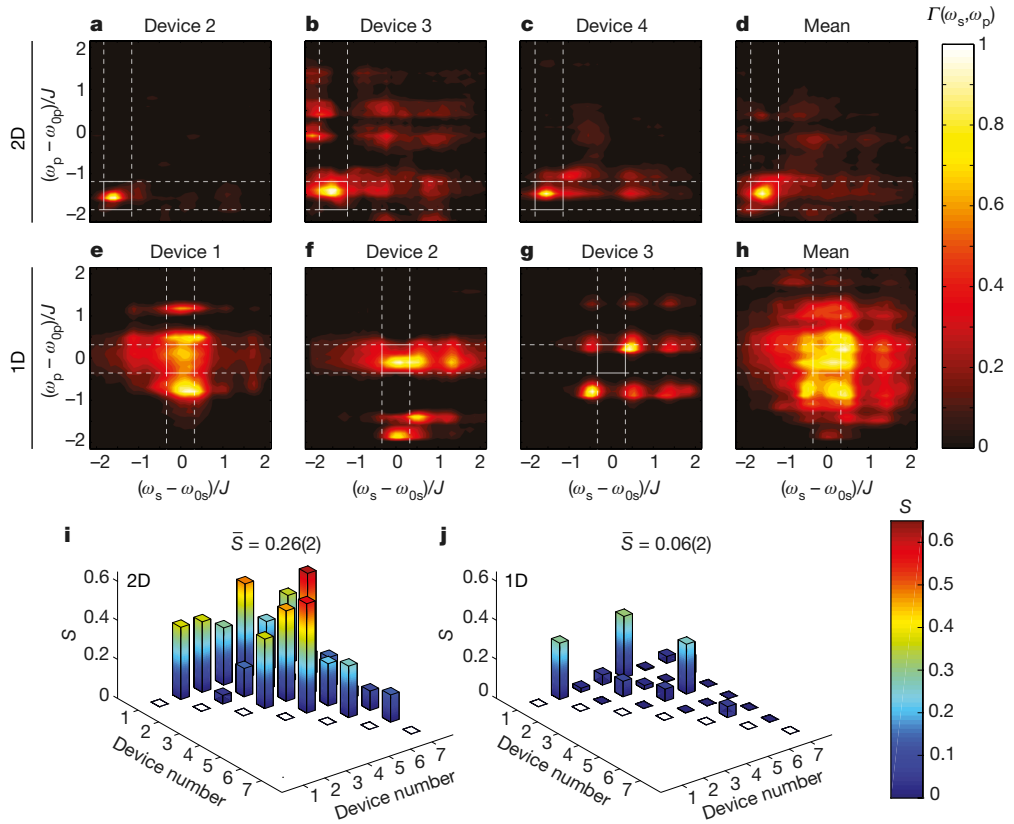


Fig. 4 | Robustness of spectral correlations between pump and signal photons. a–c, Measured $\Gamma(\omega_s, \omega_p)$, for three different 2D topological devices; **d**, the mean, measured over seven such devices (additional plots in Extended Data Fig. 1). The plots are very similar in the clockwise edge band region (marked by the solid white box), where the maximum generation of photons occurs. **e–g**, Measured $\Gamma(\omega_s, \omega_p)$, for three different 1D devices; **h**, the mean, measured over seven such devices (additional plots in Extended Data Fig. 1). There is no region of plot that is similar across all devices. The highlighted region shows the mid-band, $|\omega_\mu - \omega_{0\mu}| < 0.25J$, where transmission is maximum (see Extended Data Fig. 2). **i**, **j**, Similarity (S) of $\Gamma(\omega_s, \omega_p)$ between the edge band regions of different 2D devices (**i**) and mid-band regions of 1D devices (**j**). Because of the topological robustness, edge bands achieve a much higher similarity across devices. The error in the similarity measurement for each device pair is less than 3%, and is not shown in the figure.

Data Fig. 2) and result in device-to-device variations in the spectrum of generated photons. In the following, using measurements over many devices, we show that the topological robustness of our source manifests as a robustness in the spectrum of generated photons, outperforming the trivial 1D devices.

Figure 4a–c shows $\Gamma(\omega_s, \omega_p)$, that is, the spectrum of generated photons as a function of pump frequency, for three different 2D devices, and Fig. 4d shows the mean measured over seven devices (additional data in Extended Data Fig. 1). These devices were designed to be identical, but fabrication disorder leads to random variations in the ring resonance frequencies, coupling strengths as well as hopping phases. Nevertheless, as we saw earlier, for all devices the maximum number of photons is always generated in the clockwise edge band ($\omega_s - \omega_{0s} \approx -1.5J$), with pump frequency also in the clockwise edge band ($\omega_p - \omega_{0p} \approx -1.5J$). Therefore, in the clockwise edge band region (highlighted by dashed white lines), $\Gamma(\omega_s, \omega_p)$ is very similar for all devices. In contrast to edge bands, the spectrum of generated photons in the bulk band differs from one device to the other because it is susceptible to disorder.

Figure 4e–g show similar measurements on three different topologically trivial 1D devices, and Fig. 4h shows the mean measured over seven devices. As expected, $\Gamma(\omega_s, \omega_p)$ varies markedly from device to device, meaning that the spectral correlations between the pump and the generated photons are completely random because of the randomness in the photonic mode structure induced by fabrication disorder. Therefore, given a 1D device, the spectrum of generated photons is not known a priori for any pump frequency. To further quantify and compare the spectral correlations in 2D and 1D devices, we calculate the similarity S between $\Gamma(\omega_s, \omega_p)$ measured on two different devices (n, n'), defined as

$$S_{n,n'} = \frac{\left(\iint \sqrt{\Gamma_n \Gamma_{n'}}\, \mathrm{d}\omega_p \mathrm{d}\omega_s\right)^2}{\iint \Gamma_n \mathrm{d}\omega_p \mathrm{d}\omega_s \iint \Gamma_{n'} \mathrm{d}\omega_p \mathrm{d}\omega_s}$$

For the 2D devices, we chose the frequency integration interval to cover the clockwise edge band region $[-1.75J, -1.25J]$ which is robust against disorder and where the maximum number of photons are generated. For the 1D devices, we choose the mid-band region $[-0.25J, +0.25J]$ where the pump transmission is maximum (see Extended Data Fig. 2), and for a fair comparison with 2D devices, we choose the same bandwidth of $0.5J$ as we did for the edge region. These regions of interest are highlighted by white dashed lines in Fig. 4a–h. For the 2D system, the average similarity $\bar{S}$ across all devices is 0.26(2), whereas for the 1D system it is only 0.06(2) (Fig. 4i, j). These measurements demonstrate the advantage offered by the topological robustness of our 2D system in engineering the photonic mode structure and consequently the spectrum of generated photons.

The observed robustness and similarity in our 2D topological devices are remarkable given the fact that our system suffers from strong on-site potential disorder, comparable to the edge bandwidth[23]. To put our work into perspective, we numerically compare a 1D and topological 2D system, for slightly smaller disorder (Fig. 5). The average similarity for the 2D topological system is more than 90% and decreases only marginally as the system size increases. In comparison, for the 1D CROW, the similarity decreases rapidly with system size, approaching

© 2018 Springer Nature Limited. All rights reserved.

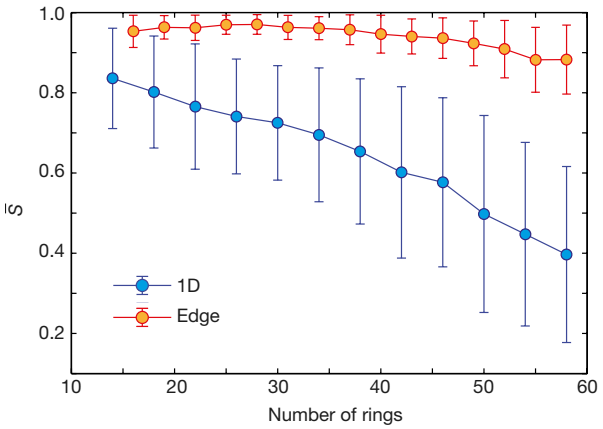


Fig. 5 | Similarity scaling as a function of device size. The simulated similarity, for moderately disordered ($V = 0.5J$) 2D and 1D systems, is shown as a function of the number of resonators travelled from the input to the output port. Because of topological protection, the 2D system achieves much higher similarity than the trivial 1D system. The results are averaged over 50 realizations of disorder. The error bars represent the standard deviation of similarity across different realizations. The solid lines are to guide the eye.

40% for 60-ring devices. The robustness of topological systems is also evident in the standard deviation of the similarity across different devices, which is considerably smaller than that of 1D CROWs. This indicates that with moderate disorder, high-visibility two-photon and multi-photon interference[11] should be possible with photons generated by different 2D topological sources.

In summary, we have demonstrated a topological source of quantum correlated photon pairs in which the spectral correlations are robust against fabrication disorder. This is a step towards on-chip, scalable sources of heralded and entangled photons with identical spectra, for applications in quantum information processing and quantum communications. Although this demonstration uses devices with relatively high propagation loss (of the order of 1 dB cm^{-1}), recent developments of ultra-low-loss photonic platforms (of the order of 10^{-3} dB cm^{-1})[31,32] could lead to orders of magnitude improvement in the source brightness. Moreover, such low-loss platforms would enable quantum-limited topological amplifiers, for which the four-wave mixing gain is required to exceed the propagation losses[26].

On a more fundamental level, we have demonstrated a robust route to manipulating the structure and vacuum fluctuations of the electromagnetic modes by using topological photonics. This could have far-reaching implications in engineering light–matter interactions in the quantum regime. We expect intriguing consequences to emerge from application of these ideas to physical phenomena such as spontaneous emission, super- and sub-radiance, and the Casimir effect.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0478-3.

Received: 5 December 2017; Accepted: 5 July 2018;
Published online 10 September 2018.

1. Shields, A. J. Semiconductor quantum light sources. *Nat. Photon.* **1**, 215–223 (2007).
2. Eisaman, M. D., Fan, J., Migdall, A. & Polyakov, S. V. Single-photon sources and detectors. *Rev. Sci. Instrum.* **82**, 071101 (2011).
3. Sharping, J. E. et al. Generation of correlated photons in nanoscale silicon waveguides. *Opt. Express* **14**, 12388–12393 (2006).
4. Clemmen, S. et al. Continuous wave photon pair generation in silicon-on-insulator waveguides and ring resonators. *Opt. Express* **17**, 16558–16570 (2009).
5. Förtsch, M. et al. A versatile source of single photons for quantum information processing. *Nat. Commun.* **4**, 1818 (2013).
6. Davanço, M. et al. Telecommunications-band heralded single photons from a silicon nanophotonic chip. *Appl. Phys. Lett.* **100**, 261104 (2012).
7. Kumar, R., Ong, J. R., Savanier, M. & Mookherjea, S. Controlling the spectrum of photons generated on a silicon nanophotonic chip. *Nat. Commun.* **5**, 5489 (2014).
8. Topolancik, J., Ilic, B. & Vollmer, F. Experimental observation of strong photon localization in disordered photonic crystal waveguides. *Phys. Rev. Lett.* **99**, 253901 (2007).
9. Mookherjea, S., Park, J. S., Yang, S.-H. & Bandaru, P. R. Localization in silicon nanophotonic slow-light waveguides. *Nat. Photon.* **2**, 90–93 (2008).
10. Sapienza, L. et al. Cavity quantum electrodynamics with anderson-localized modes. *Science* **327**, 1352–1355 (2010).
11. Spring, J. B. et al. Chip-based array of near-identical, pure, heralded single-photon sources. *Optica* **4**, 90–96 (2017).
12. Morichetti, F. et al. Travelling-wave resonant four-wave mixing breaks the limits of cavity-enhanced all-optical wavelength conversion. *Nat. Commun.* **2**, 296 (2011).
13. Lu, L., Joannopoulos, J. D. & Soljačić, M. Topological photonics. *Nat. Photon.* **8**, 821–829 (2014).
14. Wang, Z., Chong, Y., Joannopoulos, J. D. & Soljačić, M. Observation of unidirectional backscattering-immune topological electromagnetic states. *Nature* **461**, 772–775 (2009).
15. Hafezi, M., Demler, E. A., Lukin, M. D. & Taylor, J. M. Robust optical delay lines with topological protection. *Nat. Phys.* **7**, 907–912 (2011).
16. Hafezi, M., Mittal, S., Fan, J., Migdall, A. & Taylor, J. M. Imaging topological edge states in silicon photonics. *Nat. Photon.* **7**, 1001–1005 (2013).
17. Rechtsman, M. C. et al. Photonic Floquet topological insulators. *Nature* **496**, 196–200 (2013).
18. Chen, W.-J. et al. Experimental realization of photonic topological insulator in a uniaxial metacrystal waveguide. *Nat. Commun.* **5**, 5782 (2014).
19. Cheng, X. et al. Robust reconfigurable electromagnetic pathways within a photonic topological insulator. *Nat. Mater.* **15**, 542–548 (2016).
20. Kraus, Y., Lahini, Y., Ringel, Z., Verbin, M. & Zilberberg, O. Topological states and adiabatic pumping in quasicrystals. *Phys. Rev. Lett.* **109**, 106402 (2012).
21. Hafezi, M. Measuring topological invariants in photonic systems. *Phys. Rev. Lett.* **112**, 210405 (2014).
22. Mittal, S., Ganeshan, S., Fan, J., Vaezi, A. & Hafezi, M. Measurement of topological invariants in a 2D photonic system. *Nat. Photon.* **10**, 180–183 (2016).
23. Mittal, S. et al. Topologically robust transport of photons in a synthetic gauge field. *Phys. Rev. Lett.* **113**, 087403 (2014).
24. St-Jean, P. et al. Lasing in topological edge states of a one-dimensional lattice. *Nat. Photon.* **11**, 651–656 (2017).
25. Bahari, B. et al. Nonreciprocal lasing in topological cavities of arbitrary geometries. *Science* **358**, 636–640 (2017).
26. Peano, V., Houde, M., Marquardt, F. & Clerk, A. A. Topological quantum fluctuations and traveling wave amplifiers. *Phys. Rev. X* **6**, 041026 (2016).
27. Shi, T., Kimble, H. J. & Cirac, J. I. Topological phenomena in classical optical networks. *Proc. Natl Acad. Sci. USA* **114**, E8967–E8976 (2017).
28. Chen, J., Levine, Z. H., Fan, J. & Migdall, A. L. Frequency-bin entangled comb of photon pairs from a silicon-on-insulator micro-resonator. *Opt. Express* **19**, 1470–1483 (2011).
29. Ong, J. R. & Mookherjea, S. Quantum light generation on a silicon chip using waveguides and resonators. *Opt. Express* **21**, 5171–5181 (2013).
30. Yariv, A., Xu, Y., Lee, R. K. & Scherer, A. Coupled-resonator optical waveguide: a proposal and analysis. *Opt. Lett.* **24**, 711–713 (1999).
31. Bauters, J. F. et al. Ultra-low-loss high-aspect-ratio Si_3N_4 waveguides. *Opt. Express* **19**, 3163–3174 (2011).
32. Moss, D. J., Morandotti, R., Gaeta, A. L. & Lipson, M. New CMOS-compatible platforms based on silicon nitride and hydex for nonlinear optics. *Nat. Photon.* **7**, 597–607 (2013).

Acknowledgements This research was supported by AFOSR-MURI FA9550-14-1-0267, YIP-ONR, Sloan Foundation and the Physics Frontier Center at the Joint Quantum Institute. We thank V. V. Orre for help with the experimental set-up, A. Karasahin for help with the SEM, T. Huber and D. Englund for discussions and Q. Quraishi for providing the nanowire detectors.

Reviewer information *Nature* thanks V. Peano and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.M. and M.H. conceived the idea. S.M. performed the numerical simulations and the experimental measurements. E.A.G. contributed to source characterization. M.H. supervised the project. All authors contributed to analysing the data and writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0478-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Simulation of signal/idler spectra. In this section, we describe the method used to simulate the spectrum of the generated signal and idler photons in response to a strong pump field. We follow the approach described in refs [28,29]. The linear, uncoupled evolution of the pump, signal and idler fields is governed by the Hamiltonian H_L, given by equation (1) and repeated here for convenience, as

$$H_L = \sum_{m,n} \omega_{0\mu} a^{\dagger}_{m,\mu} a_{m,\mu} - J_{m,n}\left(a^{\dagger}_{m,\mu} a_{n,\mu} e^{-i\phi_{m,n}} + a^{\dagger}_{n,\mu} a_{m,\mu} e^{+i\phi_{m,n}}\right)$$

where $\mu \in \{p, s, i\}$ corresponds to the pump, signal or idler fields. The nonlinear SFWM process which couples the pump, signal and idler fields is described by the Hamiltonian (equation (2))

$$H_{NL} = \eta \sum_{m} \left(a^{\dagger}_{m,s} a^{\dagger}_{m,i} a_{m,p} a_{m,p} - a^{\dagger}_{m,p} a^{\dagger}_{m,p} a_{m,s} a_{m,i}\right)$$

Note that this Hamiltonian is local in lattice site index m. We assume that the pump field is much stronger than the signal and idler fields and therefore the evolution of the pump field is very well described by the linear Hamiltonian. However, the pump field depletes because of the intrinsic waveguide scattering losses (κ_{in}), which we include in our simulation. Using the input–output formalism and the rotating-wave approximation, we can write the coupled equations describing the steady-state pump field amplitudes, for frequency ω_p, as

$$-i\omega_p a_{m,p} = i\left[H_L, a_{m,p}\right] - \kappa_{in} a_{m,p} - (\delta_{m,I} + \delta_{m,O})\kappa_{ex} a_{m,p} - \delta_{m,I}\sqrt{2\kappa_{ex}}\, a_{in,p} \quad (3)$$

Here κ_{ex} is the coupling strength of the lattice to input/output waveguides and $a_{in,p}$ is the input pump field. The input and output waveguides are coupled to the lattice at sites indexed by I, O.

Given the pump field amplitudes calculated using equation (3), we can write the coupled equations describing the steady-state signal and idler fields amplitudes in the lattice as

$$-i\omega_\mu a_{m,\mu} = i\left[H_L + H_{NL}, a_{m,\mu}\right] - \kappa_{in} a_{m,\mu} - (\delta_{m,I} + \delta_{m,O})\kappa_{ex} a_{m,\mu} - \delta_{m,I}\sqrt{2\kappa_{ex}}\, a_{in,\mu}$$

where $\mu \in \{s, i\}$. These equations include the nonlinear FWM interaction Hamiltonian of equation (2) which couples the signal and idler fields to the pump fields. Also, for a particular choice of frequencies ω_p and ω_s, energy conservation fixes the idler frequency ω_i.

Using these coupled equations for the pump, signal and idler frequencies, we calculate their field amplitudes in the lattice. Then the signal/idler fields at the output of the lattice are calculated using the input–output formalism as

$$a_{out,\mu} = \sqrt{2\kappa_{ex}}\, a_{O,\mu}$$

where O is the index denoting the lattice output site. We can now define the spectral correlation function (SCF) $I(\omega_s, \omega_p) = |a_{out,s}|^2$. This is essentially the spectrum of generated signal photons as a function of pump frequency. Note that because of the energy conservation relation $2\omega_p = \omega_s + \omega_i$, this SCF fully characterizes the spectral correlations of the SFWM process. In other words, using $I(\omega_s, \omega_p)$, we can easily calculate $I(\omega_s, \omega_i)$, the joint-spectral intensity of the signal and idler photons.

Experimental set-up. The devices used in this experiment were fabricated using CMOS-compatible silicon-on-insulator technology. The ring resonator waveguides are approximately 510 nm in width and 220 nm in height, and at telecom wavelengths (approximately 1,550 nm) support only a single mode with transverse electric field. The coupling strength J between the resonators was measured to be 32(1) GHz and the FSR was about 1,035 GHz. The on-site disorder potential V, which is a result of the different ring resonance frequencies, was estimated to be 27.5 GHz, and the disorder on the hopping phase was 0.1. Additional details of the fabrication process and disorder characterization are available in refs [16,23].

To generate correlated photons using the SFWM process, we pumped the lattice with a telecom band, tunable, continuous-wave laser. The output of the laser was amplified using an erbium-doped fibre amplifier (EDFA), and a tunable bandpass filter was used to cut down the spontaneous emission generated during amplification. The pump was coupled to the lattice using grating couplers, with a coupling loss of about 5 dB per coupler. At the output of the lattice, tunable bandpass filters were used to remove the pump band, with a rejection exceeding 120 dB. The signal and idler photons were collected from bands separated by ±1 FSR from the pump band, respectively. To measure the spectrum of generated signal and idler photons, we used two monochromators with a frequency resolution of about 6 GHz along with two superconducting nanowire single photon detectors (PhotonSpot). The second-order correlation function measurements were done with a time-correlated single photon counting system (HydraHarp).

Source characterization. We use second-order correlation measurements to characterize our source[2,5]. The temporal correlations between signal and idler photons are analysed using the cross-correlation function $g^{(2)}_{s,i}(\tau)$ which is given as

$$g^{(2)}_{s,i}(\tau) = \frac{P_{s,i}(\tau)}{P_s P_i}$$

Here $P_{s,i}$ is the probability of detecting a signal photon at time t followed by the detection of an idler photon in the time interval $[t+\tau-\tau_c/2, t+\tau+\tau_c/2]$, and τ_c (here 50 ps) is the coincidence time-window. P_s and P_i are the probabilities of detecting individual signal or idler photons, and the product P_sP_i is the probability of detecting accidental coincidences. We observe that $g^{(2)}_{s,i}(\tau) \approx 80$ around $\tau = 0$, which implies that the generation of signal and idler photons is strongly correlated. The mean of $g^{(2)}_{s,i}(\tau)$ around $\tau = 0$ corresponds to actual coincidence counts, whereas its mean at $|\tau| \gg 0$ corresponds to accidental counts (P_sP_i). Their ratio (CAR) is commonly used as a measure of the signal-to-noise ratio of a source. We measure a maximum CAR ≈ 42 when $g^{(2)}_{s,i}$ is averaged over 300 ps (the width of the correlation peak).

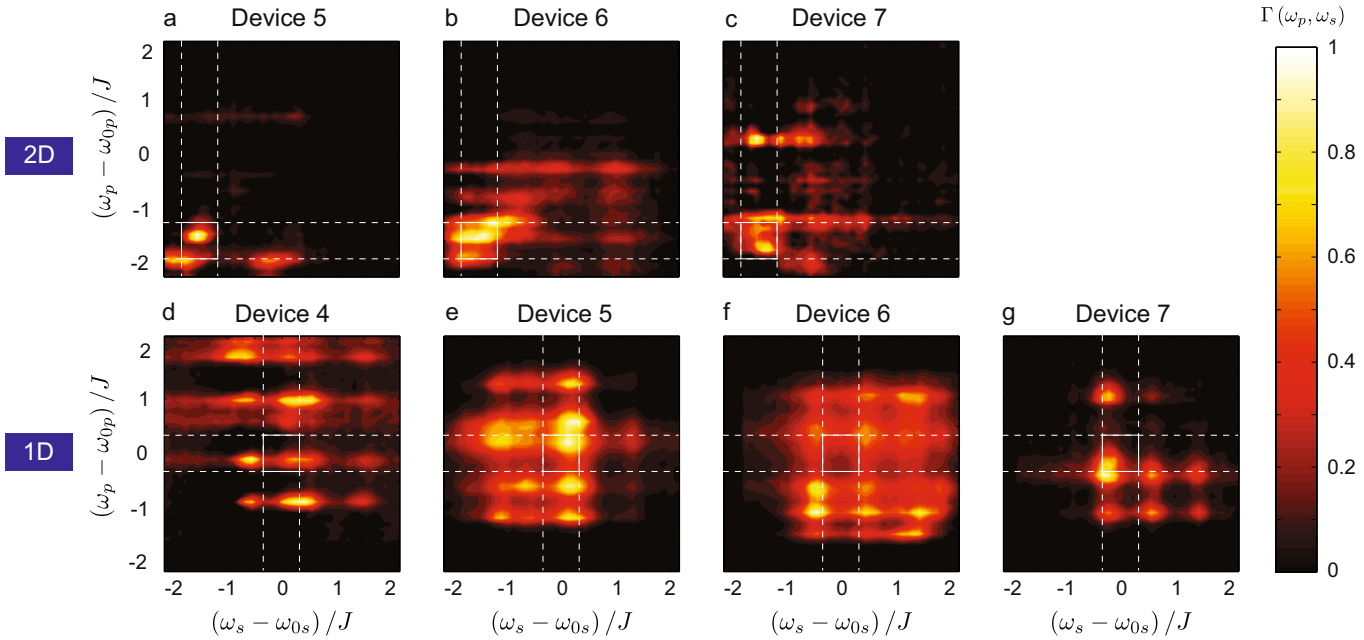
The quantum nature of a source can be demonstrated by the second-order autocorrelation function $g^{(2)}_a(\tau)$, which is a measure of antibunching of photons[2,5]. Quantum sources are distinguished by $g^{(2)}_a(0) < 1$, which suggests that the normalized probability of getting two simultaneous photons is low. In the case of correlated photon pairs, the quantum nature is revealed when we measure the conditional autocorrelation function $g^{(2)}_{a,h}(\tau)$ for signal photons heralded (conditioned) by the detection of idler photons. For this measurement, we use the Hanbury Brown–Twiss (HBT) set-up in which we place a beam-splitter in the path of signal photons, and the outputs of the beam-splitter are connected to two detectors (s_1 and s_2). The idler photons impinge on a third detector (i) which heralds the arrival of signal photons. Then the conditional autocorrelation function $g^{(2)}_{a,h}(\tau)$ for signal photons, conditioned on the detection of idler photons, is defined as

$$g^{(2)}_{a,h}(\tau) = \frac{P_{s_1,s_2,i}(\tau)}{P_{s_1,i} P_{s_2,i}}$$

Here $P_{s_1,s_2,i}(\tau)$ is the probability of detecting two heralded signal photons separated by a time τ and $P_{s_{1,2}i}$ is the probability of detecting individual heralded signal photons. These probabilities are normalized by the probability of idler (heralding) photons. Therefore, $g^{(2)}_{a,h}(0) = 0 = P_{s_1,s_2,i}(0)$ indicates that the probability of having two pairs of signal and idler photons at the same time is zero. We measure $g^{(2)}_{a,h}(0) = 0.20(8)$, which is a signature of a good source of heralded single photons.

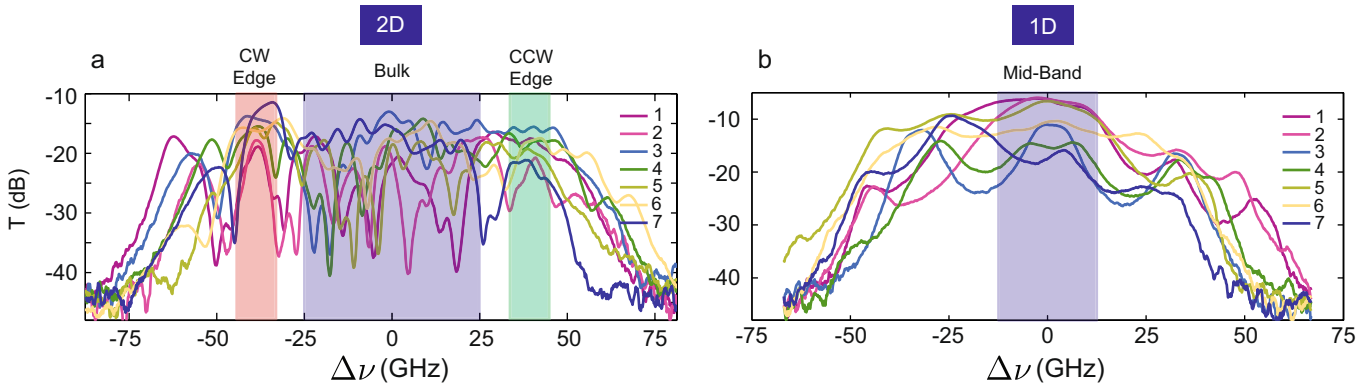
Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

© 2018 Springer Nature Limited. All rights reserved.



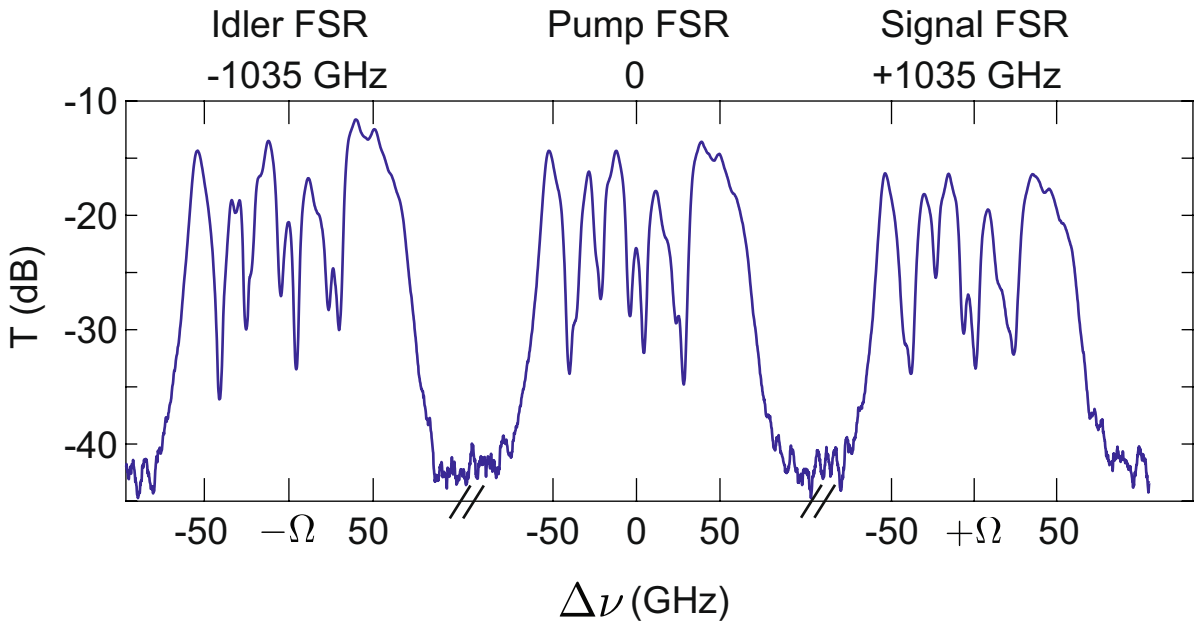
Extended Data Fig. 1 | Spectral correlations. a–c, Measured $I(\omega_s, \omega_p)$ on three different 2D devices; **d–g**, measured $I(\omega_s, \omega_p)$ on four different 1D devices, in addition to those presented in Fig. 3. Clockwise edge bands for the 2D devices and mid-band for the 1D devices are highlighted.

© 2018 Springer Nature Limited. All rights reserved.



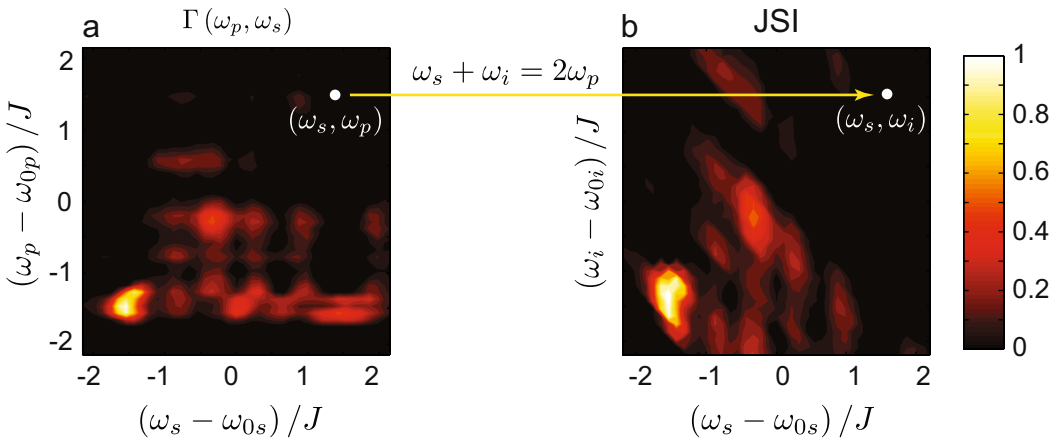
Extended Data Fig. 2 | Transmission spectra of 2D and 1D devices. **a**, **b**, Measured transmission spectra for (**a**) different 2D and (**b**) different 1D devices. The shaded regions highlight the edge and the bulk bands for the 2D system and the mid-band for the 1D system. For the 2D devices, the clockwise and the anticlockwise edge bands show reduced variations in the transmission compared with that in the bulk band. These spectra have been shifted along the frequency axis to superpose them, using an algorithm based on transmission and delay measurements, as detailed in ref. [23].

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Transmission spectrum of a 2D device across three FSRs. Measured transmission spectrum in the pump, signal and idler FSRs, corresponding to Fig. 2. $\Delta\nu$ is the frequency relative to the longitudinal mode resonance, and Ω is the FSR. The shape of the transmission spectrum in these FSRs is almost identical. The small variation in the overall transmission across bands is mainly because of the frequency response of the grating couplers.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Joint spectral intensity. a, The measured $I(\omega_s, \omega_p)$ (see Fig. 2): that is, the intensity of generated signal photons at frequency ω_s as a function of pump frequency, ω_p. Each point on this plot represents a particular ω_s and ω_p. Using energy conservation, we can calculate the corresponding idler frequency at each point as $\omega_i = 2\omega_p - \omega_s$. Therefore, we can easily rescale the *y* axis of the plot and calculate the joint-spectral intensity (JSI; see refs [11,29]) between the signal and idler frequencies, as shown in **b**. Note that this rescaling works only for a continuous-wave pump because for a pulsed pump source, the above energy conservation relation holds only up to the spectral bandwidth of the pump, signal and idler photons. Also, this measurement inherently assumes that the generated signal and idler photons are correlated. Using CAR and direct measurements of the signal and idler spectra (in Figs. 2, 3), we verified that the signal and idler photons are indeed correlated. The main advantage of such a spectral correlation measurement between the pump and the signal (or idler) photons is that it is fast and, for a continuous-wave pump, is equivalent to the JSI measurement.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0508-1

Extremely efficient terahertz high-harmonic generation in graphene by hot Dirac fermions

Hassan A. Hafez[1,2,3,7], Sergey Kovalev[4,7], Jan-Christoph Deinert[4], Zoltán Mics[2], Bertram Green[4], Nilesh Awari[4,5], Min Chen[4], Semyon Germanskiy[4], Ulf Lehnert[4], Jochen Teichert[4], Zhe Wang[4], Klaas-Jan Tielrooij[6], Zhaoyang Liu[2], Zongping Chen[2], Akimitsu Narita[2], Klaus Müllen[2], Mischa Bonn[2], Michael Gensch[4]* & Dmitry Turchinovich[1,2]*

Multiple optical harmonic generation—the multiplication of photon energy as a result of nonlinear interaction between light and matter—is a key technology in modern electronics and optoelectronics, because it allows the conversion of optical or electronic signals into signals with much higher frequency, and the generation of frequency combs. Owing to the unique electronic band structure of graphene, which features massless Dirac fermions[1–3], it has been repeatedly predicted that optical harmonic generation in graphene should be particularly efficient at the technologically important terahertz frequencies[4–6]. However, these predictions have yet to be confirmed experimentally under technologically relevant operation conditions. Here we report the generation of terahertz harmonics up to the seventh order in single-layer graphene at room temperature and under ambient conditions, driven by terahertz fields of only tens of kilovolts per centimetre, and with field conversion efficiencies in excess of 10^{-3}, 10^{-4} and 10^{-5} for the third, fifth and seventh terahertz harmonics, respectively. These conversion efficiencies are remarkably high, given that the electromagnetic interaction occurs in a single atomic layer. The key to such extremely efficient generation of terahertz high harmonics in graphene is the collective thermal response of its background Dirac electrons to the driving terahertz fields. The terahertz harmonics, generated via hot Dirac fermion dynamics, were observed directly in the time domain as electromagnetic field oscillations at these newly synthesized higher frequencies. The effective nonlinear optical coefficients of graphene for the third, fifth and seventh harmonics exceed the respective nonlinear coefficients of typical solids by 7–18 orders of magnitude[7–9]. Our results provide a direct pathway to highly efficient terahertz frequency synthesis using the present generation of graphene electronics, which operate at much lower fundamental frequencies of only a few hundreds of gigahertz.

Great demand exists in the field of modern ultrahigh-speed electronics for efficient active functional materials supporting very high, terahertz (10^{12} Hz, THz), frequencies. Since the advent of graphene, several theoretical proposals (see, for example, refs [4–6]) have predicted efficient THz frequency multiplication or harmonic generation in single-layer graphene in realistic technologically relevant scenarios, including room-temperature operation. Despite considerable efforts, the experimental verification of these predictions has so far not been achieved. Signatures of THz harmonics have recently been demonstrated[10] in a near-intrinsic epitaxial 45-layer graphene sample, but these were observable only at a cryogenic temperature. The nonlinearity mechanism in ref. [10] critically relies on inter-band THz transitions in graphene, which at a room temperature are essentially Pauli-blocked. Further, it was recently suggested that efficient THz harmonic generation in graphene is unfeasible[11] owing to the ultrafast relaxation of THz-excited electrons and holes, prohibiting coherent electronic response to driving THz fields, which is typically required for the generation of THz or optical harmonics in other solids[12,13]. Consequently, it was proposed to move away from the THz range to much higher infrared optical frequencies, where the generation of multiple optical harmonics was successfully demonstrated[11] in single-layer graphene under intense optical excitation, with peak electric field strengths reaching tens of megavolts per centimetre.

Here we show that the THz high-harmonic generation in single-layer graphene under realistic technological conditions is, in fact, an extremely efficient process, and that the THz nonlinearity of this atomically thin material greatly surpasses that of other solids. We demonstrate THz harmonic generation up to the seventh order in single-layer graphene at room temperature and under ambient conditions, using THz electric fields as weak as tens of kilovolts per centimetre, which is about three orders of magnitude lower (corresponding to peak powers one million times smaller) than that required for the observation of infrared-driven nonlinearity[11].

Previous experiments on the generation of optical harmonics in graphene critically relied on interband transitions in graphene in strong applied fields[10,11]. Here we show that the key to highly efficient THz harmonic generation in graphene is the introduction of the free background electron population, responding collectively to the driving THz field. These free background Dirac electrons, which can be easily introduced by the substrate or atmospheric doping, or by electronic gating, serve as an extremely efficient, yet highly nonlinear, intermediary energy reservoir: they facilitate and also strongly modulate the energy transfer between the THz field and graphene through a nonlinear intraband conductivity mechanism. Our experimental observations are fully reproduced by a model describing such energy transfer dynamics and the concurrent THz electromagnetic response of graphene.

Our sample is a typical monolayer graphene grown by chemical vapour deposition (CVD), deposited on a fused SiO_2 substrate (see ref. [14] and references therein). Basic characterization of the sample yielded a background free-carrier concentration of $N_c = 2.1 \times 10^{12}$ cm^{-2} and Fermi level of $E_F = 170$ meV at room temperature (see Methods and Extended Data Fig. 11).

Our experiment, schematically represented in Fig. 1a, was specifically designed for the observation of the THz harmonics in graphene in the time domain (see Methods and Extended Data Fig. 1). It was enabled by the use of TELBE, a new superconducting radiofrequency accelerator-based superradiant THz source[15]. Our measurements were performed at a repetition rate of 100 kHz and with linearly polarized THz electric fields incident normally on graphene. Narrow-band, multi-cycle quasi-monochromatic THz radiation pulses from TELBE (see Extended Data Fig. 2), with controllable peak field strength in the range of $E_f = 12$–85 kV cm^{-1} (where the subscript f indicates fundamental frequency) and full-width at half-maximum (FWHM) bandwidth below 70 GHz, were transmitted through the single-layer graphene deposited on a SiO_2 substrate and were detected in the time-domain as THz light-wave oscillations (see Methods and Extended Data Fig. 1). The THz transmission through a nominally

[1]Fakultät für Physik, Universität Duisburg-Essen, Duisburg, Germany. [2]Max Planck Institute for Polymer Research, Mainz, Germany. [3]Physics Department, Faculty of Science, Helwan University, Cairo, Egypt. [4]Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany. [5]University of Groningen, Groningen, The Netherlands. [6]ICFO—Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, Barcelona, Spain. [7]These authors contributed equally: Hassan A. Hafez, Sergey Kovalev. *e-mail: m.gensch@hzdr.de; dmitry.turchinovich@uni-due.de

© 2018 Springer Nature Limited. All rights reserved.

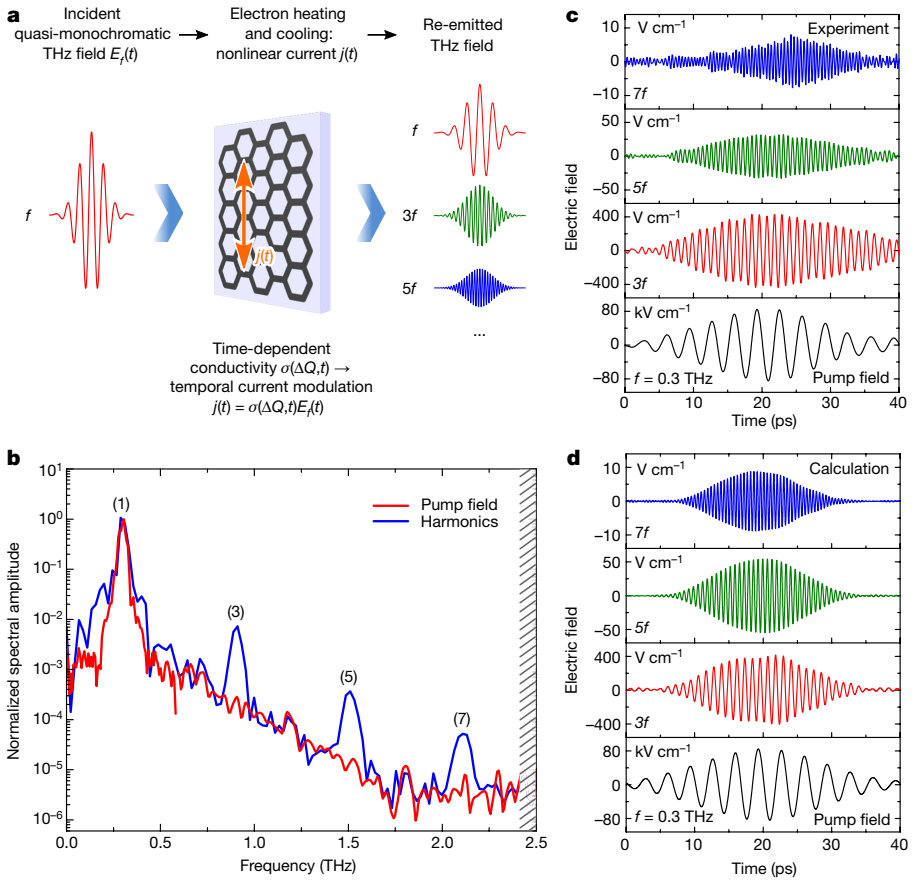


Fig. 1 | Schematic and main results of the THz high-harmonic generation. a, Schematic of the experiment: quasi-monochromatic, linearly polarized THz pump wave from the TELBE source is incident on a graphene sample, single-layer CVD-grown graphene deposited on SiO_2 substrate. The graphene has a Fermi level of $E_F = 170$ meV, free carrier density of $N_c = 2.1 \times 10^{12}$ cm^{-2} and electron momentum scattering time of $\tau = 47$ fs (all parameters correspond to full thermal equilibrium at 300 K). The peak electric field strength of the pump THz signal is controlled in the range 12–85 kV cm^{-1}, and the central frequency is varied in the range 0.3–0.68 THz. The incident THz field drives the nonlinear current in graphene, leading to re-emission at higher odd-order harmonics that appear in the spectrum of the transmitted THz signal. The THz fields are detected in the time domain directly as oscillations of electric field, using free-space electro-optic sampling with the cut-off frequency at 2.4 THz. All the experiments are carried out at room temperature and under ambient conditions. Q, electronic heat. **b**, Red line, amplitude spectrum of the incident pump THz wave at the fundamental frequency $f = 0.3$ THz with peak field strength $E_f = 85$ kV cm^{-1}, determined in the reference measurement. Blue line, the spectrum of the same THz wave transmitted through graphene on a substrate, with clearly visible generated harmonics of third, fifth and seventh order. The shaded area represents the detector cut-off. **c**, Pump wave (black line) and generated third, fifth and seventh THz harmonics for the case in **b**. The individual higher harmonics were Fourier-filtered from the complete nonlinear signal containing all the generated harmonics at once. **d**, Thermodynamic model calculation, corresponding to the measurements in **b** and **c**, using as input the experimental fundamental pump wave at frequency $f = 0.3$ THz (black line) and the basic parameters of graphene in full thermal equilibrium at 300 K.

identical bare SiO_2 substrate served as a reference, which exhibited no sign of THz nonlinearity. The cut-off frequency of our detector was 2.4 THz. All measurements were carried out at room temperature.

In Fig. 1b, c we show the typical results of our experiments. In Fig. 1b we show the amplitude spectra of the pump THz wave with a fundamental frequency of $f = 0.3$ THz and a peak electric field strength of $E_f = 85$ kV cm^{-1}, transmitted through the bare SiO_2 substrate (red line), and transmitted through the single-layer graphene on the SiO_2 substrate (blue line). Multiple THz harmonics up to the seventh order are clearly visible in the spectrum of the THz signal after interaction with graphene, with a pump-to-harmonic field conversion efficiency in excess of 10^{-3}, 10^{-4} and 10^{-5} for the third, fifth and seventh THz harmonics, respectively. Such nonlinear conversion efficiencies, especially observed at THz frequencies, are remarkable, given that the electromagnetic interaction occurs here within just one atomic layer. In Fig. 1c we demonstrate the measured THz harmonics directly in the time domain. Figure 1d shows the results of calculations based on the thermodynamic model of THz nonlinearity in graphene[14], using the experimental THz pump wave at fundamental frequency $f = 0.3$ THz as an input. The calculations in Fig. 1d capture all the key experimental features in Fig. 1c, including the peak field strength in the generated harmonics, as well as the temporal shortening of the harmonic pulses with the increase in harmonic order. The essentials of our parameter-free model[14] are explained below, and the specific calculation details are provided in the Methods.

The underlying physical mechanism of THz harmonic generation in graphene is the creation of nonlinear electrical currents resulting from the strongly nonlinear intraband THz conductivity of graphene background electrons. In electronic thermal equilibrium, the intraband conductivity of graphene σ is determined by the concentration N_c and the temperature T_e of its background electronic population. For the intraband THz excitation of graphene, N_c is constant and the conductivity σ is thus governed by T_e. If graphene is sufficiently doped, so that $|E_F| > k_B T_e$, the rise in the electronic temperature leads to a reduction of conductivity[14,16–18] (here k_B is the Boltzmann constant). The application of THz electric field E to graphene leads to generation of a THz current $j = \sigma E$, and hence to deposition of energy into the electronic population (the power absorption coefficient of a material is directly proportional to its conductivity). One of the remarkable features of Dirac electrons in graphene is their ability to thermalize internally—that is, exchange

© 2018 Springer Nature Limited. All rights reserved.

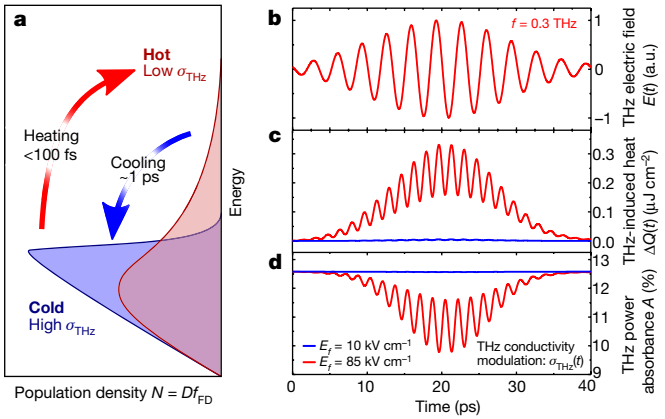


Fig. 2 | Illustration of the mechanism of THz harmonic generation in graphene, based on calculations with the thermodynamic model of intraband nonlinear THz conductivity of graphene. a–d, Temporal asymmetry of electron heating and cooling rates in graphene (**a**) leads to a situation where the application of a THz driving field (**b**) to the graphene electron population results in electronic heat accumulation $\Delta Q(t)$ (**c**) and hence in the temporal modulation (**d**) of the THz absorbance $A(t)$ and of the conductivity of graphene $\sigma(t)$. In these calculations the driving field (**b**) oscillates at the fundamental frequency $f = 0.3$ THz. Blue and red lines in **c** and **d** correspond to driving field peak amplitudes of 10 kV cm^{-1} and 85 kV cm^{-1}, respectively. The field strength of 10 kV cm^{-1} represents the regime of small-signal nonlinearity (no substantial modification of the sample properties on the timescale of the interaction with the THz wave), whereas the field strength of 85 kV cm^{-1} leads to considerable electron heating and thus to a substantial decrease in the absorption of graphene at the peak of interaction.

their kinetic energy and establish a common electronic temperature within the entire background electron population—on an ultrafast timescale below 100 femtoseconds[14,19–24], which is quasi-instantaneous as compared with the picosecond period of the THz field oscillation. Therefore, through such an ultrafast internal electronic thermalization, the excess energy deposited into graphene background electrons by the THz current is quasi-instantaneously converted into collective kinetic energy of the entire background electron population, that is into electronic heat, leading to a rise in T_e and to a concomitant decrease of σ. As a result of these interdependencies, the THz conductivity of graphene depends strongly nonlinearly on the driving THz field[14,17,25,26]: the stronger the field, the smaller the conductivity becomes.

The subsequent energy relaxation (cooling) of the hot electron population in graphene occurs through phonon emission on a timescale of a few picoseconds (see, for example, refs [20,25])—that is, with THz rate (see Methods). This phonon emission completes the energy conversion process in the THz excitation of graphene, in which the background population of Dirac electrons serves as a nonlinear intermediary in the transfer of energy from the absorbed THz field to the lattice of graphene.

The processes of heating and cooling of background Dirac electron population, leading to the reduction and recovery of the intraband THz conductivity in graphene, respectively, are schematically shown in Fig. 2a. Here $N(E) = D(E)f_{FD}(E,T)$ is the energy-dependent population density function for graphene background electrons, a product of the density of states D and the Fermi–Dirac distribution f_{FD}. Now we arrive at a situation in which the THz conductivity of graphene is not only nonlinear with the driving THz field (as facilitated by the direct conversion of absorbed THz energy into electronic heat) but also exhibits THz-rate recovery dynamics provided by the electron cooling rate (transfer of electronic heat to the lattice). The double interplay of quasi-instantaneous electron heating and THz-rate electron cooling in THz-driven graphene leads to highly nonlinear temporal modulation of the THz absorption in graphene, and hence of the THz conductivity, as illustrated in the calculations shown in Fig. 2b–d (see Methods for details). Consequently, a nonlinear current modulated by nonlinear time-dependent conductivity is generated in graphene in response to the driving THz field $E(t)$. In the case of a monochromatic THz driving field $E_f(t)$ oscillating at a frequency f, the resulting nonlinear current in graphene, given its centrosymmetry, will contain contributions at odd-order harmonics f, $3f$, $5f$, $7f$, …, thus leading to electromagnetic re-emission at these harmonic frequencies (as illustrated schematically in Fig. 1a) and resulting in the appearance of these harmonics in the spectrum of the transmitted THz signal. In a centrosymmetric medium, the even-order harmonics cancel out, and thus cannot be generated[7]. Using the above thermodynamic model[14], the nonlinear propagation of THz signals through graphene can be readily calculated, only using experimentally determined parameters as input: the incident THz waveform at the fundamental frequency f; the conductivity parameters of the graphene sample measured in the linear regime in full thermal equilibrium at $T = 300$ K; the quasi-instantaneous electron heating (see discussion above and refs [19–24]); and picosecond-timescale cooling through phonon emission as measured in ref. [20]. We also note that no adjustable parameters are used in our model (see Methods).

Now we determine the harmonic generation efficiency and estimate the effective THz nonlinear optical susceptibility of graphene from the analysis of experimentally measured time-domain THz signals, such as those shown in Fig. 1c. In Fig. 3a, we show the dependency of the amplitude of the generated harmonics fields on that of the driving field. Here, to increase the dynamic range in the characterization of the third and fifth harmonic generation processes, our THz detection scheme was optimized to detect the harmonic signal only in the spectral window around 2 THz, whereas the driving fundamental frequency f of the THz source was varied as follows: $f = 0.68$ THz was used to generate the third harmonic at $3f = 2.04$ THz; $f = 0.37$ THz was used to generate the fifth harmonic at $5f = 1.85$ THz; and $f = 0.3$ THz was used to generate the seventh harmonic at $7f = 2.1$ THz. Depending on the order of the generated harmonic, the peak harmonic field strength was in the range 3–111 V cm^{-1}, whereas the pumping field strength was varied in the range 12–85 kV cm^{-1}. The efficiency parameters of the THz harmonic generation in this work are summarized in Table 1.

In the same Fig. 3a, as black solid lines, we present the results of the calculated pump-to-harmonic conversion efficiencies, using the parameter-free thermodynamic model[14] as explained above. The agreement between the experimental data and the parameter-free calculation is quantitative. The harmonic generation efficiency, shown in Fig. 3a, saturates with the pump field increase, as caused by the dissipative nature of the interaction between the THz field and the graphene electrons. In Fig. 2c, d we show the calculated temporal evolution of the THz field-induced electronic heat $\Delta Q(t)$ and the instantaneous THz power absorbance $A(t)$ of our graphene sample, driven by the THz field at the fundamental frequency $f = 0.3$ THz shown in Fig. 2b, and with the field strength at the lower and upper limits of our experiment: 10 kV cm^{-1} and 85 kV cm^{-1}, respectively. At the weaker driving field of 10 kV cm^{-1}, the evolution of both $\Delta Q(t)$ and $A(t)$ is very small. Here, the nonlinear interaction conditions between the THz field and graphene remain approximately constant over the entire interaction period, thus manifesting the small-signal nonlinearity regime. On the other hand, as the driving THz field increases by about an order of magnitude to 85 kV cm^{-1}, the effect of the THz field on the electronic system of graphene becomes very strong. At the peak of the interaction, the electronic heat $\Delta Q(t)$ reaches the value of 0.33 μJ cm^{-2}, leading to a reduction of the THz power absorbance of graphene by approximately 25%. This considerable absorption saturation naturally limits the energy transfer rate from the driving THz field to the electronic system, subsequently leading to the observed saturable behaviour of THz nonlinearity in graphene. The marked difference between the conditions of THz–graphene interaction for the cases of $E_f = 10$ kV cm^{-1} and $E_f = 85$ kV cm^{-1} in Fig. 2c, d illustrates the extremely nonlinear nature of the thermodynamic response of the Dirac electrons in graphene to the THz fields.

Importantly, at lower driving fields below approximately 20 kV cm^{-1}, both the data and the calculation of harmonic generation efficiency

© 2018 Springer Nature Limited. All rights reserved.

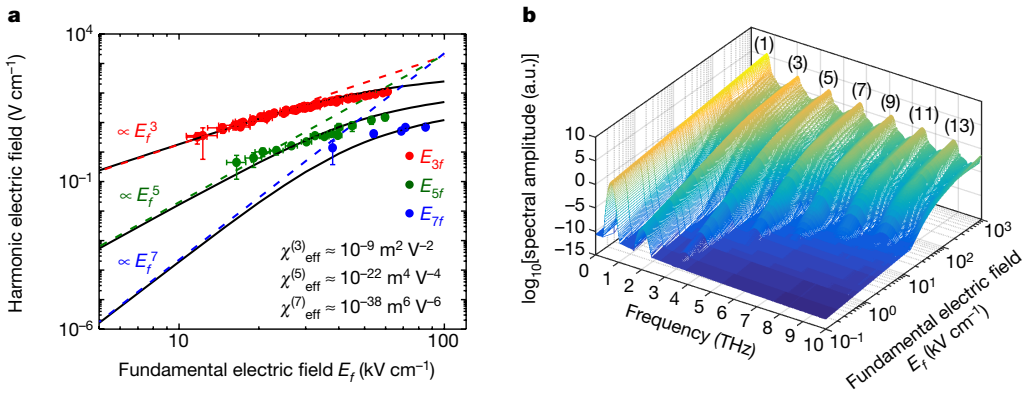


Fig. 3 | Efficiency of THz high-harmonic generation. a, Dependence of the generated harmonic field strength on the fundamental pump field strength. For the third, fifth and seventh harmonics, the fundamental field frequency was $f = 0.68$ THz, $f = 0.37$ THz and $f = 0.3$ THz, respectively, leading to corresponding higher harmonics of $3f = 2.04$ THz, $5f = 1.85$ THz and $7f = 2.1$ THz, respectively. Symbols represent data; black lines, parameter-free calculations using the thermodynamic model; dashed lines, power of 3 (red), power of 5 (green) and power of 7 (blue) fits corresponding to the small-signal nonlinearity regime below the driving field of about 20 kV cm^{-1}, which allows estimation of the third-, fifth- and seventh-order THz effective nonlinear susceptibilities of graphene, $\chi^{(3)}_{eff}$, $\chi^{(5)}_{eff}$ and $\chi^{(7)}_{eff}$, as indicated in the figure. The error bars for the pump fields are defined by the width of the chosen pulse intensity distribution. The error bars for the harmonic fields are derived from the sum of the standard deviations from multiple measurements and the background noise. **b**, Calculation of higher-order harmonic generation in graphene using the thermodynamic mechanism: amplitude spectra of the THz field transmitted through the graphene sample. The calculation was performed with sample parameters as used in this work ($N_c = 2.1 \times 10^{12}$ cm^{-2} and $E_F = 170$ meV), using the model incident THz field at $f = 0.68$ THz with peak strength in the range of 100 V cm^{-1} to 1 MV cm^{-1}. Harmonics up to 13th order appear in the spectrum within the 10^6 dynamic range with respect to the transmitted fundamental.

in Fig. 3a reduce to the phenomenological power-law dependencies $E_{mf} \propto E_f^m$, where m is the harmonic order (see dashed lines in Fig. 3a), manifesting the small-signal nonlinearity regime. In this regime, the nonlinear conversion can be described within the framework of effective odd-order nonlinear optical coefficients $\chi^{(m)}_{eff}$, precisely as expected for the nonlinear response in a centrosymmetric medium. From the purely power-law fits to the field conversion efficiencies (dashed lines in Fig. 3a), we determine the effective THz third-, fifth- and seventh-order nonlinear coefficients of graphene as $\chi^{(3)}_{eff} \approx 10^{-9}$ m^2 V^{-2}, $\chi^{(5)}_{eff} \approx 10^{-22}$ m^4 V^{-4} and $\chi^{(7)}_{eff} \approx 10^{-38}$ m^6 V^{-6}, respectively (see Methods for details). These values of THz nonlinear optical coefficients of graphene, determined by us at room temperature and under ambient conditions in a conventional CVD-grown single-layer graphene, exceed the corresponding nonlinearities of a typical solid by 7–18 orders of magnitude[7–9].The third-order nonlinearity obtained in this work exceeds the corresponding values measured in single-layer graphene in the optical spectral range by 6–10 orders of magnitude[27–29] (we note, however, that the theoretical prediction of stronger optical-range response exists in the literature[30]); and the value reported in the infrared for the 50-layer graphene sample under cryogenic conditions and in strong magnetic field by at least one order of magnitude[31]. The fifth- and seventh-order THz nonlinear coefficients of graphene have not previously been reported in the literature, to the best of our knowledge.

We note here that the thermodynamic model of graphene conductivity does not contain any a priori defined effective nonlinear-optical coefficients of graphene, but rather describes the temporal evolution of the THz field-induced electronic heat and, hence, the instantaneous conductivity of graphene during its interaction with the incident THz electromagnetic wave of arbitrary shape. Very good agreement between our parameter-free calculation and the entirety of our data lends credence to our interpretation of the nature of the observed THz harmonic generation in graphene.

We estimate the limits of applicability of the purely thermodynamic picture of nonlinear THz response in doped graphene as the onset of coherent Bloch oscillations (see, for example, ref. [13]), which in graphene will occur at a pump field strength of 1–10 MV cm^{-1} (see Methods). We now illustrate the potential for higher-order THz harmonic generation in graphene using the thermodynamic mechanism, for typical experimental and technological scenarios. With the same material parameters of graphene as used in this work, we apply a model quasi-monochromatic waveform at fundamental $f = 0.68$ THz with varying peak field strengths in the range 100 V cm^{-1} to 1 MV cm^{-1}. This range comprises the linear regime of THz fields (of the order of a few kV cm^{-1}) and the regime of typical high-speed transistor channel fields (of the order of 100 kV cm^{-1})[32], and extends to the TELBE operation regime under projected design parameters of 1 MV cm^{-1} (ref. [15]). In Fig. 3b we show the corresponding calculated amplitude spectra of the transmitted THz waveform as a function of the peak electric field strength of the incident fundamental wave. As the driving electric field strength increases, higher-order harmonics keep appearing in the spectrum, with the 13th harmonic falling within the 10^6 dynamic range with respect to the transmitted fundamental, at a field strength for the driving signal of 1 MV cm^{-1}.

The very high THz nonlinear coefficients of graphene, measured in this work, suggest straightforward applications in ultrahigh-frequency (opto-)electronics. A driving electric field of tens of kilovolts per centimetre, as used here, is about one order of magnitude smaller than the typical channel field in ultra-high speed transistors[32]. Therefore, our results provide a direct pathway to purely electronic THz frequency synthesis within the present generation of graphene transistors operating at fundamental frequencies of a few hundred gigahertz[32]. Further, the observed nonlinear THz response of a single layer graphene can be made scalable, for example, by creating stratified structures with graphene-coated interfaces, or graphene-loaded waveguides or cavities.

Table 1 | Harmonic generation efficiencies and nonlinear coefficients of graphene

Generated harmonic order, m	Range of pump field strength, E_f (kV cm^{-1})	Range of harmonic field strength, E_{mf} (V cm^{-1})	Maximum field conversion efficiency, E_{mf}/E_f	Effective nonlinear coefficient, $\chi^{(m)}_{eff}$
3	12–61	3–111	2×10^{-3}	~10^{-9} m^2 V^{-2}
5	17–60	0.4–15	2.5×10^{-4}	~10^{-22} m^4 V^{-4}
7	38–85	1–7	8×10^{-5}	~10^{-38} m^6 V^{-6}

Ranges of pump and harmonic peak electric field strength, maximum field conversion efficiency and the THz nonlinear coefficients of graphene of third, fifth and seventh orders, established in the small-signal nonlinearity regime, are shown.

© 2018 Springer Nature Limited. All rights reserved.

Finally, we note that our results on efficient THz harmonic generation in graphene by hot Dirac fermions could be potentially generalized to other representatives of the new class of Dirac materials, such as topological insulators, or Weyl and Dirac semimetals[33].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0508-1.

Received: 11 December 2017; Accepted: 28 June 2018;
Published online 10 September 2018.

1. Novoselov, K. S. et al. Two-dimensional gas of massless Dirac fermions in graphene. *Nature* **438**, 197–200 (2005).
2. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81**, 109–162 (2009).
3. Das Sarma, S., Adam, S., Hwang, E. H. & Rossi, E. Electronic transport in two-dimensional graphene. *Rev. Mod. Phys.* **83**, 407–470 (2011).
4. Mikhailov, S. A. Non-linear graphene optics for terahertz applications. *Microelectronics J.* **40**, 712–715 (2009).
5. Al-Naib, I., Sipe, J. E. & Dignam, M. M. High harmonic generation in undoped graphene: interplay of inter- and intraband dynamics. *Phys. Rev. B* **90**, 245423 (2014).
6. Al-Naib, I., Poschmann, M. & Dignam, M. M. Optimizing third-harmonic generation at terahertz frequencies in graphene. *Phys. Rev. B* **91**, 205407 (2015).
7. Boyd, R. W. *Nonlinear Optics* 3rd edn (Academic, New York, 2008).
8. Reyna, A. S. & de Araújo, C. B. High-order optical nonlinearities in plasmonic nanocomposites—a review. *Adv. Opt. Photonics* **9**, 720–724 (2017).
9. Reshef, O. et al. Beyond the perturbative description of the nonlinear optical response of low-index materials. *Opt. Lett.* **42**, 3225–3228 (2017).
10. Bowlan, P., Martinez-Moreno, E., Reimann, K., Elsaesser, T. & Woerner, M. Ultrafast terahertz response of multilayer graphene in the nonperturbative regime. *Phys. Rev. B* **89**, 041408 (2014).
11. Yoshikawa, N., Tamaya, T. & Tanaka, K. High-harmonic generation in graphene enhanced by elliptically polarized light excitation. *Science* **356**, 736–738 (2017).
12. Zaks, B., Liu, R. B. & Sherwin, M. S. Experimental observation of electron–hole recollisions. *Nature* **483**, 580–583 (2012).
13. Schubert, O. et al. Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. *Nat. Photon.* **8**, 119–123 (2014).
14. Mics, Z. et al. Thermodynamic picture of ultrafast charge transport in graphene. *Nat. Commun.* **6**, 7655 (2015).
15. Green, B. et al. High-field high-repetition-rate sources for the coherent THz control of matter. *Sci. Rep.* **6**, 22256 (2016).
16. Winnerl, S. et al. Ultrafast processes in graphene: from fundamental manybody interactions to device applications. *Ann. Phys.* **529**, 1700022 (2017).
17. Jadidi, M. M. et al. Nonlinear terahertz absorption of graphene plasmons. *Nano Lett.* **16**, 2734–2738 (2016).
18. Hafez, H. A. et al. Nonlinear terahertz field-induced carrier dynamics in photoexcited epitaxial monolayer graphene. *Phys. Rev. B* **91**, 035422 (2015).
19. Breusing, M. et al. Ultrafast nonequilibrium carrier dynamics in a single graphene layer. *Phys. Rev. B* **83**, 153410 (2011).
20. Gierz, I. et al. Snapshots of non-equilibrium Dirac carrier distributions in graphene. *Nat. Mater.* **12**, 1119–1124 (2013).
21. Gierz, I. et al. Tracking primary thermalization events in graphene with photoemission at extreme time scales. *Phys. Rev. Lett.* **115**, 086803 (2015).
22. Winzer, T. & Malić, E. Impact of Auger processes on carrier dynamics in graphene. *Phys. Rev. B* **85**, 241404 (2012).
23. Song, J. C. W., Tielrooij, K. J., Koppens, F. H. L. & Levitov, L. S. Photoexcited carrier dynamics and impact-excitation cascade in graphene. *Phys. Rev. B* **87**, 155429 (2013).
24. Tomadin, A., Brida, D., Cerullo, G., Ferrari, A. C. & Polini, M. Nonequilibrium dynamics of photoexcited electrons in graphene: collinear scattering, Auger processes, and the impact of screening. *Phys. Rev. B* **88**, 035430 (2013).
25. Winnerl, S. et al. Carrier relaxation in epitaxial graphene photoexcited near the Dirac point. *Phys. Rev. Lett.* **107**, 237401 (2011).
26. Hwang, H. Y. et al. Nonlinear THz conductivity dynamics in p-type CVD-grown graphene. *J. Phys. Chem. B* **117**, 15819–15824 (2013).
27. Hendry, E., Hale, P. J., Moger, J., Savchenko, A. K. & Mikhailov, S. A. Coherent nonlinear optical response of graphene. *Phys. Rev. Lett.* **105**, 097401 (2010).
28. Hong, S. et al. Optical third-harmonic generation in graphene. *Phys. Rev. X* **3**, 021014 (2013).
29. Kumar, N. et al. Third harmonic generation in graphene and few-layer graphite films. *Phys. Rev. B* **87**, 121406 (2013).
30. Cheng, J. L., Vermeulen, N. & Sipe, J. E. Third-order nonlinearity of graphene: effects of phenomenological relaxation and finite temperature. *Phys. Rev. B* **91**, 235320 (2015).
31. König-Otto, J. C. et al. Four-wave mixing in Landau-quantized graphene. *Nano Lett.* **17**, 2184–2188 (2017).
32. Schwierz, F. Graphene transistors: status, prospects, and problems. *Proc. IEEE* **101**, 1567–1584 (2013).
33. Wehling, T. O., Black-Schaffer, A. M. & Balatsky, A. V. Dirac materials. *Adv. Phys.* **63**, 1–76 (2014).

Acknowledgements D.T. acknowledges financial support from the Deutsche Forschungsgemeinschaft (SFB 1242 'Non-Equilibrium Dynamics of Condensed Matter in the Time Domain', TP B08), European Commission (EU Career Integration Grant EU CIG 334324 LIGHTER) and the Max Planck Society. M.G. and B.G. acknowledge support from the European Cluster of Advanced Laser Light Sources (EUCALL) project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no 654220. K.J.T. acknowledges support through the Mineco Young Investigator Grant (FIS2014-59639-JIN). J.T. and U.L. acknowledge support from the EuCARD-2 project, which has received funding from the European Commission under grant agreement No 312453. We thank J. Lægsgaard, K. Krewer, E. Unger, W. Zhang, T. V. A. G. de Oliveira and M. Mittendorff for discussions. We thank the ELBE team for the operation of the TELBE facility.

Author contributions D.T. and M.G. conceived and supervised the project. H.A.H., S.K., J.-C.D., Z.M., B.G., N.A., M.C., S.G., Z.W., D.T. and M.G. performed the nonlinear THz spectroscopy measurements and evaluated the experimental data. H.A.H. and D.T. performed the modelling, with contributions from Z.M. and K.-J.T. D.T., M.G. and M.B. interpreted the results, with contributions from all co-authors. Z.L., Z.C., A.N. and K.M. manufactured the samples. H.A.H. and D.T. characterized the linear THz properties of the graphene/fused silica samples. U.L. and J.T. provided for the special mode of high bunch charge operation of TELBE using the SRF photoinjector that enabled the observation of the seventh harmonic. M.B. initiated and supported the THz studies on graphene at MPI-P. D.T. and M.G. wrote the manuscript, with contributions from M.B., K.-J.T., S.K. and H.A.H. All co-authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0508-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.G. and D.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Basic experimental set-up. The experiments were performed at the TELBE THz facility, which provides tunable, narrow-band (bandwidth of 20%, FWHM), spectrally dense THz pulses in the frequency range between 0.1 THz and 1.2 THz with adjustable repetition rates up to the few 100-kHz regime. The multicycle THz pulses are generated in an undulator through the process of superradiance from ultra-short relativistic electron bunches accelerated and compressed in a quasi-continuous-wave superconducting radiofrequency accelerator[15]. In the standard mode of operation, a thermionic injector serves as the electron source and was used in most experiments. A recent upgrade now also allows the use of a specially developed photoinjector that provides higher electron bunch charges. This was used to perform the experiments at 0.3 THz central frequency, in particular enabling the observation of the seventh THz harmonic.

In these specific experiments, linearly polarized THz pulses with central frequencies 0.3 THz, 0.37 THz and 0.68 THz were used separately as input fundamental driving field. The input THz beam was focused on the graphene sample at normal incidence, and the transmitted (that is, re-emitted) THz waves were detected behind the sample (see Extended Data Fig. 1). All measurements were carried out at room temperature and under ambient atmosphere. The higher harmonic background was suppressed by appropriate shaping of the longitudinal electron bunch form and by additional narrow bandpass filters[34] before the sample, creating highly monochromatic pump fields with a bandwidth of roughly 10% (FWHM) (Extended Data Fig. 2). On transmission through the sample, the THz signal containing the remainder of the fundamental field, as well as the generated harmonics, was again sent through a set of bandpass filters with the transmission band centred at the generated harmonic to be detected. The resulting THz field was measured directly in the time domain by free-space electro-optic sampling (FEOS) in a 1.9-mm-thick ZnTe crystal[35], with the peak THz field strength not exceeding about 0.6 kV cm^{-1}. For FEOS THz detection, we used gating pulses of 100 fs duration, central wavelength 805 nm, from a commercial Ti:sapphire laser system (https://cohrcdn.azureedge.net/assets/pdf/RegA-Data-Sheet.pdf), which is synchronized and timed to the TELBE source[15]. The uncertainty in the timing between the THz pulses generated by the superradiant THz source and the probe laser pulses was less than 30 fs (FWHM)[36]. The specific choice of the 1.9-mm-thick ZnTe crystal enables sensitive THz field sampling only in the frequency range below 2.4 THz, which motivated our choice of the fundamental frequencies of $f = 0.68$ THz, $f = 0.37$ THz and $f = 0.3$ THz for harmonic generation at $3f = 2.04$ THz, $5f = 1.85$ THz and $7f = 2.1$ THz, respectively.

To achieve the optimal dynamic range required for measuring the field dependence of these different harmonics shown in Fig. 3a of the main text, specifically chosen bandpass filter combinations were used in the individual experimental configurations described in more detail later. Combined with the advanced pulse-resolved data acquisition scheme[36], dynamic ranges of better than 10^3 in the time-domain detection of the seventh and third harmonic fields, and 10^9 in the case of the fifth harmonic measurement, were achieved.

Pulse-resolved determination of THz intensities and fields. Depending on the tuning of the TELBE THz source, intensities can fluctuate between 10% and 50% with respect to the average value. The intensity of each THz pulse is recorded using a fast pyro detector with a bandwidth of 1 MHz in front of the first bandpass filter (see Extended Data Fig. 1). The THz pulses are thereafter grouped and evaluated in separate ranges around a certain pyro detector level. The width of these ranges defines the error bar in the pump fields in Fig. 3a. The absolute values for the intensity and the corresponding THz field are derived from a careful cross-calibration carried out before the experiment. For this, the intensity seen by the pyro detector was compared to that of the FEOS measurement at different pyro detector levels. The THz intensity was dimmed in front of the pyro detector and in front of the ZnTe crystal to avoid nonlinearities in both detection schemes. The peak fields at the sample position were determined from a combination of the beam profile, as measured by a commercial pyroelectric camera of the type Ophir Pyrocam III (https://www.ophiropt.com/laser-measurement/beam-profilers/products/Beam-Profiling/Camera-Profiling-with-BeamGage/Pyrocam-IIIHR), the THz pulse shape and intensity, as measured by the combination of FEOS and pyro detector, and the THz power, as determined by a commercial calibrated THz power-meter of the type Ophir 3A-P-THz (https://www.ophiropt.com/laser-measurement/laser-power-energy-meters/products/Laser-Thermal-Power-Sensors/High-Sensitivity-Thermal-Laser-Sensors/3A-P-THz)[37,38].

Simultaneous detection of the third, fifth and seventh harmonic. In this experiment, a fundamental frequency of 0.3 THz was chosen for generation and detection of the seventh harmonic at 2.1 THz. The repetition rate of the TELBE THz source was 100 kHz; the femtosecond laser had a repetition rate of 200 kHz. The particular configuration of bandpass filters, as shown in Extended Data Fig. 3, allowed not only the detection of the seventh harmonic but also simultaneous observation of the third and fifth harmonics at 0.9 THz and 1.5 THz (see Fig. 1). Two 0.3-THz bandpass filters were used to suppress the harmonic background of the undulator, leading to a maximum field strength of 85 kV cm^{-1} at the sample. The resulting suppression factor in intensity is 10^{-8} for the third, 10^{-6} for the fifth and 10^{-4} for the seventh harmonic. The filtered THz beam was focused on the sample by an off-axis parabolic mirror in combination with an additional Teflon lens, which yielded a THz spot size of 560 μm (FWHM). A 2.1-THz bandpass filter was used behind the sample to enable optimal sensitivity for the seventh harmonic in the FEOS measurement, while still allowing a detectable portion of the lower harmonics and the fundamental to pass through.

The as-measured higher harmonic generation (HHG) spectrum from graphene is shown in Extended Data Fig. 4 (red curve). It clearly shows the fundamental and harmonics up to the seventh order. The raw signal originating from the bare SiO_2 substrate as a reference is shown in black. To determine the corresponding electric fields quantitatively, the transmission function of the 2.1-THz filter and the frequency-dependent sensitivity of the ZnTe detection crystal must be included. The former has been measured and is displayed in Extended Data Fig. 4 as the grey curve, showing the substantial suppression of frequencies below 1.9 THz. The spectra presented in Fig. 1b were obtained by deconvoluting the response functions of both the ZnTe detection crystal and the 2.1-THz filter from the measured FEOS signals (see discussion in the next section on reconstructing the harmonic fields from the measured FEOS signals and Extended Data Fig. 9). In order not to overly amplify the noise above 2.2 THz in the graphene sample and reference spectra, the ZnTe sensitivity curve was assumed to be flat for values above 2.2 THz in this case. This does not affect either the overall peak shape in Fig. 1 or the peak field values that were used to calculate the efficiencies of HHG (see Fig. 3a and Table 1).

Separate detection of the third and fifth harmonic at 2.04 THz and 1.85 THz. The repetition rate of the TELBE THz source was set to 101 kHz; the femtosecond laser used for FEOS detection had a repetition rate of 202 kHz. To explicitly measure and quantify (see Fig. 3a) the electric field strength of the third harmonic generation (THG) and fifth harmonic generation (FHG) signals, we chose the fundamental frequencies of 0.68 THz and 0.37 THz, respectively. Sets of narrow-bandpass filters[34] were tailored correspondingly. These filters suppress the higher harmonic background to less than 10^{-7} for 0.68 THz and to less than 10^{-11} for 0.37 THz in intensity. Thereby, we reached an optimal dynamic range in the FEOS detection of the harmonic field, despite the fact that the harmonic field is weaker than the fundamental field by a factor of 10^{-3} (THG) or 10^{-4} (FHG). Beam diameters (FWHM) on the sample surface were 1 mm for 0.68 THz and 0.6 mm for 0.37 THz. In the latter case, an additional Teflon lens was used to optimize the spot size.

In Extended Data Fig. 5a, we show a schematic of the beampath in the THG experiment. Here a single 1.93-THz bandpass filter[34] was introduced after the sample to attenuate the fundamental at 0.68 THz, while allowing transmission of the THG signal at 2.04 THz. The spectral transmission function of the 1.93-THz filter is shown in Extended Data Fig. 6a. Extended Data Fig. 6b shows the as-measured amplitude spectra of the detected FEOS signals: the reference (pump field transmitted through a bare SiO_2 substrate) and that from the sample (pump field transmitted through graphene on a SiO_2 substrate). The pump field contains a small spurious contribution at the THG frequency (black line in Extended Data Fig. 6b), which is, however, an order of magnitude smaller than the THG generated in graphene (red line in Extended Data Fig. 6b).

In the case of the THG experiment, the leaking field at the fundamental frequency $f = 0.68$ THz measured through the bare SiO_2 substrate (geometry as shown in Extended Data Fig. 5a) serves as a representative for the fundamental THz excitation field in Extended Data Fig. 10 and was used as a reference input pulse in the theoretical model.

In the FHG experiment, we used the beampath as schematically shown in Extended Data Fig. 5b. To ensure greater sensitivity for the detection of the fifth harmonic field, the number of bandpass filters before and after the sample was doubled for all fields except for the three highest field values shown in Fig. 3a, for which the 0.4-THz filter was removed. Extended Data Fig. 7 shows the as-measured amplitude spectra of the detected reference and sample FEOS signals. The transmission at the fundamental frequency is strongly suppressed, and no spurious contribution of the pump field at the FHG frequency is observed.

Because the signal of the 0.37-THz fundamental in this configuration was small, it was (in contrast to the THG experiment) not suitable to derive the waveform of the fundamental THz pulse from the leaking fundamental. Instead, the waveform was measured with an additional experimental configuration shown in Extended Data Fig. 5c. This waveform was also used as a reference input pulse for the theoretical calculation of the nonlinear response of graphene.

Reconstruction of the harmonic fields from the measured FEOS signals. The complex-valued (that is, containing both amplitude and phase) field transmission functions of all optical elements in the THz beampath of our experiments (sample substrate, bandpass filters), as well as the FEOS detector (THz Fresnel insertion losses, and the frequency-dependent acceptance function of the FEOS ZnTe crystal which is described in refs [39,40]) have been characterized, as summarized below.

© 2018 Springer Nature Limited. All rights reserved.

As a result, rigorous reconstruction of the measured FEOS signals back to the actual, calibrated THz electric field transients at all frequencies of interest (fundamental and higher harmonics) was possible. Our reconstruction from the FEOS signals back to the actual fields follows the same general protocol reported in ref. [41], while taking into account all elements in the THz beampath. Here we exemplify the field reconstruction from the measured EOS using the THz THG experiment with $f = 0.68$ THz $\rightarrow 3f = 2.04$ THz.

Extended Data Fig. 6a shows the amplitude transmission spectrum of the 1.93 THz bandpass filter, while Extended Data Fig. 8 shows the phase and thereby time difference added to the signals at all frequencies of interest owing to the propagation through this optical element. A THz time-domain spectroscopy measurement of the SiO_2 substrate yielded values of frequency-dependent power absorption coefficient of 7.6 cm^{-1} (2.04 THz), 6.6 cm^{-1} (1.85 THz), 2.3 cm^{-1} (0.68 THz) and 1.3 cm^{-1} (0.37 THz), and a spectrally flat refractive index of 1.98, in good agreement with the literature[42].

In Extended Data Fig. 9a, we show the frequency-dependent transmission response function (amplitude transmission and phase shift) of the bare SiO_2 substrate, while in Extended Data Fig. 9b we show the acceptance function (amplitude and phase shift)[39] of our FEOS unit: 1.9-mm-thick ZnTe crystal, gated by the 100-fs pulses at 805-nm central wavelength.

Taking all the above elements used in the THz transmission and detection in our experiment, we are now able to rigorously reconstruct the as-measured THz FEOS signals back into the actual THz-field waveforms[41], as shown in Extended Data Fig. 10 (example of THG measurement with $f = 0.68$ THz $\rightarrow 3f = 2.04$ THz).

Extended Data Fig. 10a shows the measured FEOS signals transmitted through the bare substrate (black pulse), through the graphene/substrate sample (red pulse), both normalized to the peak of the substrate pulse, and the difference between the two normalized FEOS signals (blue pulse). The difference pulse (the blue one) clearly shows oscillations at the THG frequency of 2.04 THz. Extended Data Fig. 10b shows the corresponding reconstructed fields (the fields at the point just after the graphene–substrate interface) after compensating for the field losses and deconvoluting the response functions, described above, of the optical elements after the graphene–substrate interface. We note that deconvoluting the phase shift induced by these optical elements after the graphene film from the FEOS signals in Extended Data Fig. 10a results in a time-shift of about 17.8 ps to the reconstructed fields in Extended Data Fig. 10b. Further, in the reconstructed fields in Extended Data Fig. 10b, the fundamental field is upscaled to its original values before the bandpass filter attenuation after the sample, and in turn dominates over the THG. Thus, the oscillation at the THG frequency included in the difference field pulse (blue pulse) in Extended Data Fig. 10b is now less pronounced than that in Extended Data Fig. 10a. Now, if we spectrally remove the fundamental signal at 0.68 THz from the field difference pulse shown in Extended Data Fig. 10b, we arrive at the THG field waveform (Extended Data Fig. 10c).

Sample preparation and characterization. Our sample is a typical CVD-grown monolayer graphene transferred onto a fused silica substrate[14,43,44]. The graphene quality was tested by Raman spectroscopy, which showed a single layer identity with a large-scale homogeneity (see Extended Data Fig. 11a for the Raman spectra of the sample). The characteristic graphene 2D and G peaks[45] are observed at 2703 cm^{-1} and 1591 cm^{-1}, respectively. The Gaussian shape of the 2D band with FWHM of about 47 cm^{-1}, along with the relative ratio of the 2D peak to the G peak, denotes single layer identity of the graphene sample. The very weak D band at 1,300 cm^{-1} is indicative of a low defect density.

The linear conductivity of our graphene sample was characterized by THz time-domain spectroscopy using a ZnTe crystal as a THz emitter pumped by 800-nm laser pulses with a pulse width of 40 fs and a repetition rate of 1 kHz. The experimental data revealed Drude-type conductivity[14,46–49], providing the Fermi level energy of 170 meV and electron momentum scattering time of 47 fs (see Extended Data Fig. 11b for the linear THz characterization of the sample).

We note here that our graphene sample is, in fact, p-type, as is typical for CVD-grown material deposited on fused SiO_2 substrate in the presence of ambient atmosphere (see for example ref. [50]). However, owing to the symmetry of the band structure for Dirac electrons and holes in graphene, for simplicity we depict the Fermi level in the conduction rather than in the valence band.

Calculations. The (nonlinear) complex-valued THz field transmission from the free space through graphene into the substrate is described by the Tinkham equation[51]

$$\tilde{E}_t(\omega) = \frac{2}{n_s + 1 + Z_0\tilde{\sigma}(\omega)}\tilde{E}_{in}(\omega) \tag{1}$$

where n_s is the refractive index of the substrate, Z_0 is the free-space impedance and $\tilde{\sigma}(\omega)$ is the complex THz conductivity of the conductive film, and $\tilde{E}_{in}(\omega)$ is the complex-valued THz incident field. The conductivity $\tilde{\sigma}(\omega)$ is nonlinear on the driving THz field[14,18,25,26,52,53]; see also the discussion below.

The intraband conductivity of conduction-band electrons of graphene at THz frequencies $\tilde{\sigma}(\omega)$ can generally be described by the solution of the Boltzmann equation[3,14,54]:

$$\tilde{\sigma}(\omega) = -\frac{e^2 v_F^2}{2}\int_0^\infty D(E)\frac{\tau(E)}{1 - i\omega\tau(E)}\frac{\partial f_{FD}(E, \mu, T_e)}{\partial E}dE \tag{2}$$

where $v_F = 1 \times 10^6$ m s^{-1} is the Fermi velocity of the relativistic Dirac fermions in graphene, $\tau(E)$ is the energy-dependent electron scattering time, $D(E) = 2|E|/\pi(\hbar v_F)^2$ is the density of states of the graphene energy bands, and $f_{FD}(E, \mu, T_e) = \{\exp[(E-\mu)/(k_B T_e)] + 1\}^{-1}$ is the Fermi–Dirac distribution function for a chemical potential μ and electron temperature T_e.

The dominating electron momentum scattering mechanism in CVD-grown graphene deposited on fused silica substrate is through Coulomb interactions with charged impurities (long-range Coulomb scattering)[3,14,55–58], which is defined by linear dependence of scattering time on energy $\tau(E) = \gamma E$ (see also ref. [59]). The proportionality constant $\gamma = \tau_0/E_F = 47$ fs/(170 meV) can be obtained from the linear response described above (see also ref. [14] and Extended Data Fig. 11b). In terms of the THz-field-induced heat density ΔQ, one can express $\tau(\Delta Q) = \gamma(E_F + \Delta Q/N_c) = \tau_0 + \gamma\Delta Q/N_c$, where $N_c = \int_0^\infty D(E) f_{FD}(E, \mu, T_e)dE$.

To compute the time-dependent THz-field-induced heat ΔQ accumulated in the electronic system of graphene during the (nonlinear) interaction with the driving THz wave, split-step time-domain calculations, as common in nonlinear optics, must be performed (see for example ref. [60]). For this, the inverse Fourier transforms of equation (1) with the conductivity $\tilde{\sigma}(\omega)$ described by equation (2) must be taken at every time step during numerical propagation. To speed up the calculations, a linear parameterization of equation (2) has been used, under the conditions of energy conservation (all absorbed THz energy adds to ΔQ) and free-carrier density conservation ($N_c = \int_0^\infty D(E) f_{FD}(E, \mu, T_e)dE = \text{constant}$), yielding the following expression

$$\tilde{\sigma}(\omega) \approx \frac{(ev_F)^2\gamma N_c}{1 - i\omega\tau(\Delta Q)} \tag{3}$$

Now, the time-domain expression of the THz field transmission of graphene, by taking the inverse Fourier transform of equation (1) with the conductivity $\tilde{\sigma}(\omega)$ described by equation (3), results in a time-dependent response function

$$\mathcal{T}(t) = \frac{2\delta(t)}{n_s + 1} - \frac{2Z_0(ev_F)^2\gamma N_c}{[\tau_0 + \gamma\Delta Q(t)/N_c](n_s + 1)^2} \times \exp\left\{-\frac{Z_0(ev_F)^2\gamma N_c + n_s + 1}{[\tau_0 + \gamma\Delta Q(t)/N_c](n_s + 1)}t\right\} \tag{4}$$

where $\delta(t)$ is the Dirac delta function. The convolution of this function with the temporal THz driving field $E_{in}(t)$ yields the corresponding temporal field transmitted through the graphene film into the SiO_2 substrate:

$$E_t(t) = \mathcal{T}(t) * E_{in}(t) \tag{5}$$

where the asterisk denotes the convolution.

To accommodate for the electron cooling due to phonon emission, we use the following cooling rate as directly measured in ref. [20]

$$R(t) = 0.001\exp\left(-\frac{t}{\tau_1}\right) + 0.46\exp\left(-\frac{t}{\tau_2}\right) + 0.539\exp\left(-\frac{t}{\tau_3}\right) \tag{6}$$

with time constants $\tau_1 = 13$ fs, $\tau_2 = 80$ fs and $\tau_3 = 0.86$ ps. This relaxation process leads to a recovery towards the initial linear conductivity of the graphene after the incident THz pulse ceases. The calculation of the THz-induced heat is thus based on the following convolution:

$$\Delta Q(t) = \frac{1}{2}\varepsilon_0 c R(t) * \{E_{in}^2(t) - n_s E_t^2(t) - [E_t(t) - E_{in}(t)]^2\} \tag{7}$$

where ε_0 is the vacuum permittivity.

As a result, the time-dependent induced electronic heat given by equation (7) leads to the temporal modulation of conductivity and results in the nonlinear response function of graphene described by equation (4), which in turn leads to harmonic generation.

The nonlinear THz refractive index of graphene. The dielectric function of the graphene film can be expressed by[61,62]

$$\varepsilon_r = \varepsilon_\infty + \frac{i\tilde{\sigma}}{d\omega\varepsilon_0} \tag{8}$$

where $d = 0.3$ nm is the thickness of the graphene film and $\varepsilon_\infty \approx 1$ (refs [61,62]). We calculate the real part of the refractive index using $n = \text{Re}\{\sqrt{\varepsilon_r}\}$. Substituting the

© 2018 Springer Nature Limited. All rights reserved.

linear Drude conductivity shown in Extended Data Fig. 11b into equation (8), we first obtain the linear (field-independent) refractive index (for fields less than 1 kV cm^{-1}). When the graphene film interacts with higher fields, the conductivity decreases as explained above, leading to a reduction in the refractive index (Extended Data Fig. 12). In Extended Data Fig. 12a, we show the calculated field-dependent THz index when the graphene film is pumped by THz fields at a fundamental frequency of 0.3 THz. In Extended Data Fig. 12b, we show the THz field dependence of the nonlinear refractive index at the harmonics $3f = 2.04$ THz generated by the $f = 0.68$ THz pump, $5f = 1.85$ THz generated by the $f = 0.37$ THz pump, and $7f = 2.1$ THz generated by the $f = 0.3$ THz pump. From this field dependence of the refractive index, we estimated the third-, fifth- and seventh-order susceptibilities $\chi^{(3)}$, $\chi^{(5)}$ and $\chi^{(7)}$, respectively, as explained in the next section.

Extraction of nonlinear coefficients of graphene. The conservative, low-boundary estimate of the third-, fifth- and seventh-order THz nonlinear optical coefficients of graphene, based on the experimentally measured field conversion efficiencies in small-signal regime (power-law fits to the data and the results of thermodynamic model in Fig. 3a) is done as follows. Ignoring both the pump field depletion (given approximately 90% field transmission through graphene in the most absorptive limit of small driving THz fields), and the phase mismatch in $f \rightarrow mf$ conversion processes (negligible within one atomic layer), the conversion from the pump to the harmonic field is defined as $|E_{mf}| = \gamma L |E_f|^m$ with $\gamma = k_m[\pi f/(cn_{mf})]\chi^{(m)}_{\mathrm{eff}}$ (refs [7,63]). Here $m = (3,5,7)$ is the harmonic order; E_f and E_{mf} are peak electric field strengths of the incident fundamental and generated mth harmonic field, respectively; $k_m = (3/4, 5/16, 7/64)$ is a numerical coefficient for harmonics $m = 3$, 5 and 7, respectively; $L = 0.3$ nm is the thickness of graphene layer[64]; c is the speed of light in vacuum; and n_{3f}, n_{5f} and n_{7f} represent the refractive indices of graphene at the harmonic frequencies, which at the target harmonic frequency around 2 THz is chosen to be about 10 (see below) . From the measured THG with $f = 0.68$ THz $\rightarrow 3f = 2.04$ THz, the field conversion coefficient of $\gamma L = |E_{3f}|/|E_f|^3 = 1.79 \times 10^{-16}$ m^2 V^{-2}, and an estimate of the effective third-order THz nonlinear susceptibility of graphene is $\chi^{(3)}_{\mathrm{eff}} \approx 1.7 \times 10^{-9}$ m^2 V^{-2}. For the measured FHG with $f = 0.37$ THz $\rightarrow 5f = 1.85$ THz, the field conversion coefficient $\gamma L = |E_{5f}|/|E_f|^5 = 2.68 \times 10^{-30}$ m^4 V^{-4} and the effective fifth-order THz nonlinear susceptibility of graphene is estimated as $\chi^{(5)}_{\mathrm{eff}} \approx 1.2 \times 10^{-22}$ m^4 V^{-4}. In the same fashion, for the seventh harmonic generation (7HG) process with $f = 0.3$ THz $\rightarrow 7f = 2.1$ THz, the field conversion coefficient $\gamma L = |E_{7f}|/|E_f|^7 = 1.8 \times 10^{-46}$ m^6 V^{-6} and the estimate of the effective seventh-order THz nonlinear susceptibility of graphene is $\chi^{(7)}_{\mathrm{eff}} \approx 1.74 \times 10^{-38}$ m^6 V^{-6}.

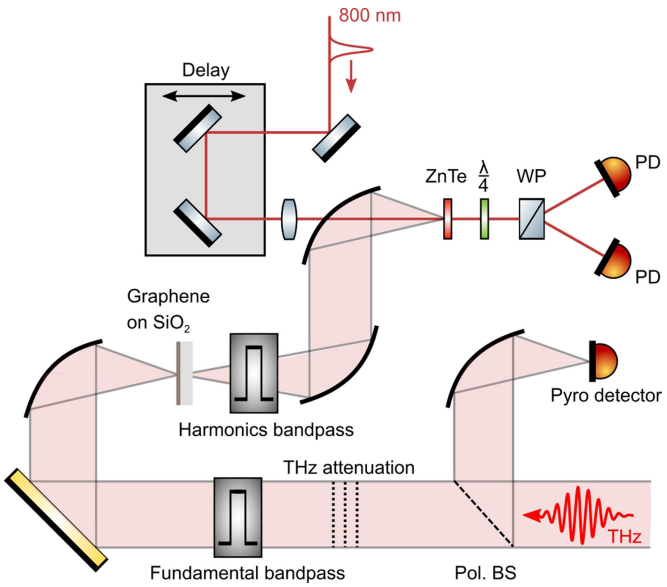
We note that the value for the refractive index of graphene of $n_{3f,5f,7f} \approx 10$ (at the frequency around 2 THz, and at high electron temperature) used in our calculations is a conservative order-of-magnitude estimate, as the actual values can be as high as $n \approx 20$–80 (Extended Data Fig. 12b), which will increase the values of the retrieved nonlinear coefficients of graphene even further.

Limits of the thermodynamic model of THz response of graphene in terms of driving THz field strength. The electron response, following the heating–cooling dynamics of electron population as described in this paper, will dominate the THz intraband nonlinearity of graphene until the regime of coherent Bloch oscillations will be reached. In this regime (see, for example, refs [13,65]), the conduction band electron reaches the Brillouin zone boundary in the time faster than the typical electron momentum scattering (Drude) time, providing for the coherent nonlinear current (Bloch) oscillations in the k space of the Brillouin zone between the $k = \pi/a$ and $-k = -\pi/a$ points, where a is the lattice constant. We use the acceleration theorem $\mathrm{d}k/\mathrm{d}t = -eE/\hbar$, which can be further simplified to $\Delta k/\tau = -eE/\hbar$, where $\Delta k = \pi/a$ is the Brillouin zone half-width, $a = 0.14$ nm is the lattice constant of graphene, and τ is the electron momentum relaxation time. For the typical electron momentum relaxation time in the range $\tau = 10$–100 fs (see, for example, ref. [3]) we arrive at the electric field strength range $E = 1$–10 MV cm^{-1}. This field strength thus sets the estimated upper validity limit for our model.

Data and code availability. The datasets generated and analysed during this study, and the corresponding computer codes, are available from the corresponding authors on reasonable request.

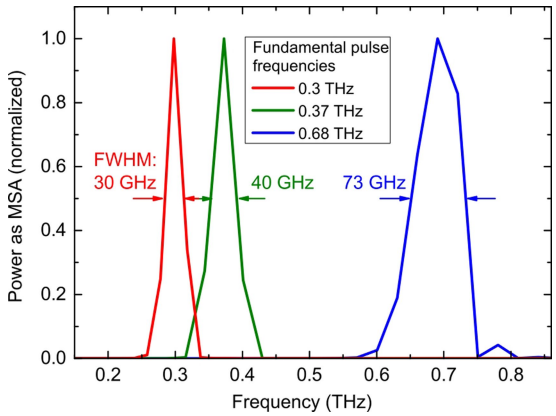
34. Kuznetsov, S. A., Astafyev, M. A., Gelfand, A. V. & Arzhannikov, A. V. Microstructured frequency selective quasi-optical components for submillimeter-wave applications. In *44th European Microwave Conference (EuMC)*, 881 (IEEE, Rome, 2014).
35. Wu, Q., Litz, M. & Zhang, X. C. Broadband detection capability of ZnTe electro-optic field detectors. *Appl. Phys. Lett.* **68**, 2924–2926 (1996).
36. Kovalev, S. et al. Probing ultra-fast processes with high dynamic range at 4th-generation light sources: arrival time and intensity binning at unprecedented repetition rates. *Struct. Dyn.* **4**, 024301 (2017).
37. Steiger, A., Kehrt, M., Monte, C. & Müller, R. Traceable terahertz power measurement from 1 THz to 5 THz. *Opt. Express* **21**, 14466–14473 (2013).
38. Müller, R. et al. Novel detectors for traceable THz power measurements. *J. Infrared Millim. Terahertz Waves* **35**, 659–670 (2014).
39. Gallot, G., Zhang, J., Mcgowan, R. W., Jeon, T. & Grischkowsky, D. Measurements of the THz absorption and dispersion of ZnTe and their relevance to the electro-optic detection of THz radiation. *Appl. Phys. Lett.* **74**, 3450 (1999).
40. Wu, Q. & Zhang, X. Ultrafast electro-optic field sensors. *Appl. Phys. Lett.* **68**, 1604 (1996).
41. Kampfrath, T. et al. Terahertz spin current pulses controlled by magnetic heterostructures. *Nat. Nanotech.* **8**, 256–260 (2013).
42. Parker, T. J. & Ford, J. E. The optical constants of pure far-infrared. *Infrared Phys.* **18**, 215–219 (1978).
43. Li, X. et al. Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science* **324**, 1312–1314 (2009).
44. Hao, Y. et al. The role of surface oxygen in the growth of large single-crystal graphene on copper. *Science* **342**, 720–723 (2013).
45. Ferrari, A. C. et al. Raman spectrum of graphene and graphene layers. *Phys. Rev. Lett.* **97**, 187401 (2006).
46. Mak, K. F., Ju, L., Wang, F. & Heinz, T. F. Optical spectroscopy of graphene: from the far infrared to the ultraviolet. *Solid State Commun.* **152**, 1341–1349 (2012).
47. Jnawali, G., Rao, Y., Yan, H. & Heinz, T. F. Observation of a transient decrease in terahertz conductivity of single-layer graphene induced by ultrafast optical excitation. *Nano Lett.* **13**, 524–530 (2013).
48. Dawlaty, J. M. et al. Measurement of the optical absorption spectra of epitaxial graphene from terahertz to visible. *Appl. Phys. Lett.* **93**, 131905 (2008).
49. Frenzel, A. J. et al. Observation of suppressed terahertz absorption in photoexcited graphene. *Appl. Phys. Lett.* **102**, 113111 (2013).
50. Wang, H. I. et al. Reversible photochemical control of doping levels in supported graphene. *J. Phys. Chem. C* **121**, 4083–4091 (2017).
51. Glover, R. E. & Tinkham, M. Conductivity of superconducting films for photon energies between 0.3 and 40 kT_c. *Phys. Rev.* **108**, 243–256 (1957).
52. Suess, R. J. et al. Role of transient reflection in graphene nonlinear infrared optics. *ACS Photonics* **3**, 1069–1075 (2016).
53. Hafez, H. A. et al. Intense terahertz field effects on photoexcited carrier dynamics in gated graphene. *Appl. Phys. Lett.* **107**, 251903 (2015).
54. Marder, M. P. in *Condensed Matter Physics*, 483–493 (Wiley, Hoboken, NJ, 2010).
55. Tielrooij, K. J. et al. Photoexcitation cascade and multiple hot-carrier generation in graphene. *Nat. Phys.* **9**, 248–252 (2013).
56. Hwang, E. H. & Das Sarma, S. Acoustic phonon scattering limited carrier mobility in two-dimensional extrinsic graphene. *Phys. Rev. B* **77**, 115449 (2008).
57. Hwang, E. H. & Das Sarma, S. Screening-induced temperature-dependent transport in two-dimensional graphene. *Phys. Rev. B* **79**, 165404 (2009).
58. Perebeinos, V. & Avouris, P. Inelastic scattering and current saturation in graphene. *Phys. Rev. B* **81**, 195442 (2010).
59. Lundstrom, M. *Fundamentals of Carrier Transport* 2nd edn (Cambridge Univ. Press, New York, 2000).
60. Liu, X., Laegsgaard, J. & Turchinovich, D. Self-stabilization of a mode-locked femtosecond fiber laser using a photonic bandgap fiber. *Opt. Lett.* **35**, 913–915 (2010).
61. Lin, I. et al. Terahertz optical properties of multilayer graphene: experimental observation of strong dependence on stacking arrangements and misorientation angles. *Phys. Rev. B* **86**, 235446 (2012).
62. Gosciniak, J. & Tan, D. T. H. Theoretical investigation of graphene-based photonic modulators. *Sci. Rep.* **3**, 1897 (2013).
63. Saltiel, S. M., Sukhorukov, A. A. & Kivshar, Y. S. in *Progress in Optics* Vol. 47 (ed. Wolf, E.) 1–73 (Elsevier, Amsterdam, 2005).
64. Shearer, C. J., Slattery, A. D., Stapleton, A. J., Shapter, J. G. & Gibson, C. T. Accurate thickness measurement of graphene. *Nanotechnology* **27**, 125704 (2016).
65. Hohenleutner, M. et al. Real-time observation of interfering crystal electrons in high-harmonic generation. *Nature* **523**, 572–575 (2015).

© 2018 Springer Nature Limited. All rights reserved.



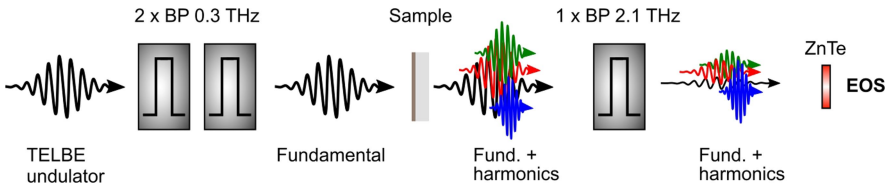
Extended Data Fig. 1 | Experimental set-up. Tunable multicycle THz pulses (red) from the undulator of the TELBE facility[2] are used to irradiate the graphene sample. 100-fs pulses from a Ti:sapphire laser system (brown) are used to probe the transmitted and emitted THz pulses by free-space electro-optic sampling. PD, photodiode; BS, beamsplitter; Pol., polarizer.

© 2018 Springer Nature Limited. All rights reserved.



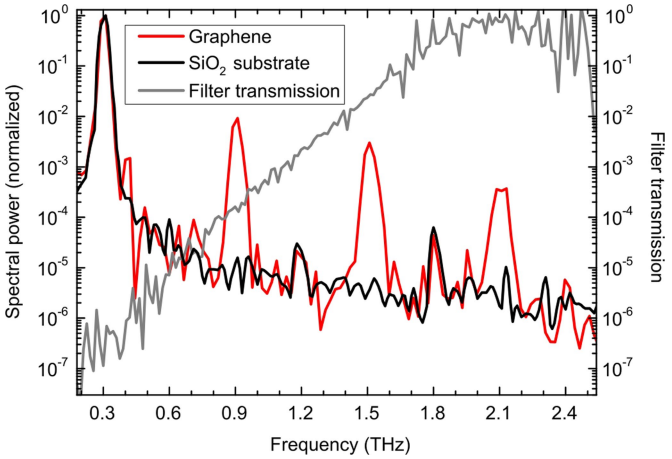
Extended Data Fig. 2 | Fundamental frequencies after bandpass filtering. The bandwidths were determined from Gaussian fits to the spectra. MSA, mean square amplitude.

© 2018 Springer Nature Limited. All rights reserved.



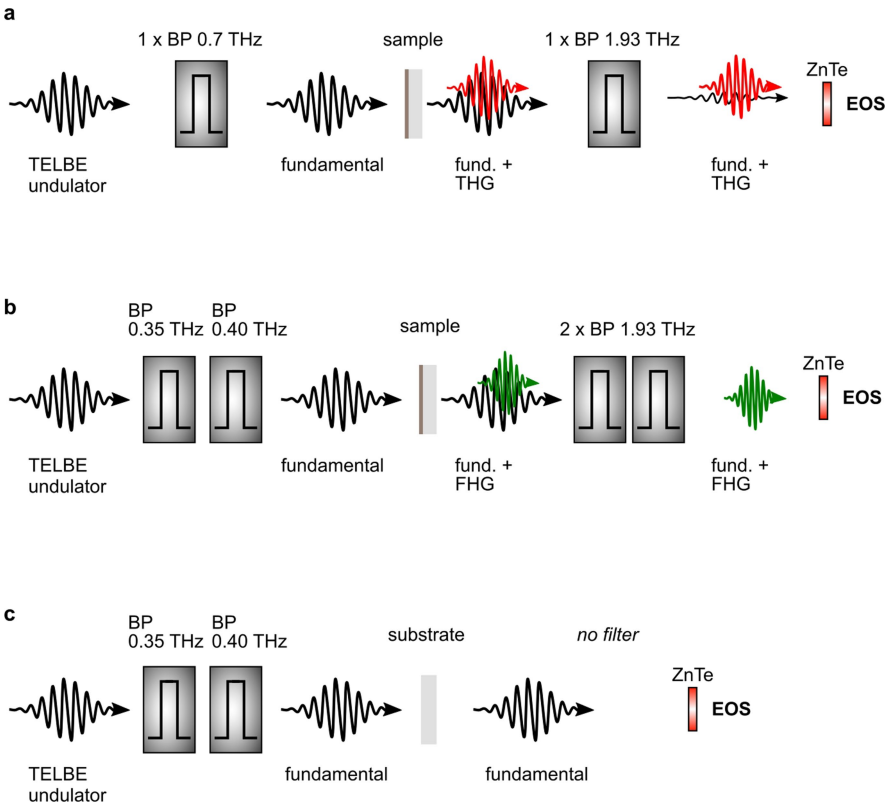
Extended Data Fig. 3 | Scheme of the set-up for detection of multiple harmonics up to the seventh order. Two 0.3-THz bandpass filters (BP) are used to suppress the undulator harmonic background. A single 2.1-THz bandpass filter after the sample attenuates the fundamental, third and fifth harmonics to an extent that they can still be detected by the EOS set-up.

© 2018 Springer Nature Limited. All rights reserved.



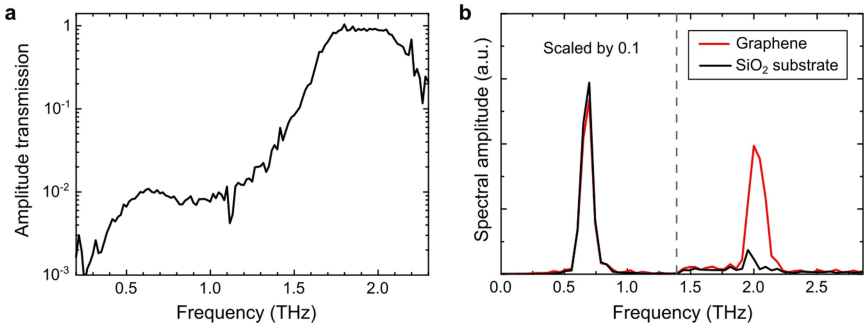
Extended Data Fig. 4 | HHG signal from graphene, reference signal from the SiO_2 substrate and filter function of the 2.1-THz bandpass. The red curve shows the as-measured HHG spectrum of the graphene sample. The black curve shows the reference spectrum taken from the bare SiO_2 substrate. The measured transmission function of the 2.1-THz bandpass filter is also shown (grey line).

© 2018 Springer Nature Limited. All rights reserved.



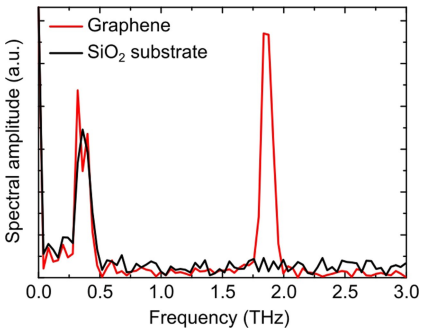
Extended Data Fig. 5 | Schemes of the experimental configurations to determine the electric fields of the fundamental, THG and FHG pulses. Measurements were performed with graphene/SiO_2 and with the bare SiO_2 substrate. **a**, Set-up for the THG experiment used to measure the fundamental and harmonic simultaneously. **b**, Set-up to measure the harmonic in the FHG experiment. Two filters were used before the sample and two after the sample, to optimize the signal-to-noise ratio. **c**, Set-up to determine the electric fields for the fundamental in the FHG experiment.

© 2018 Springer Nature Limited. All rights reserved.



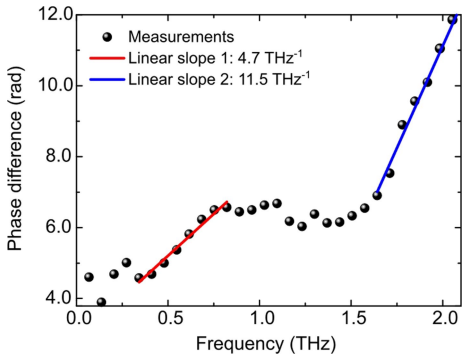
Extended Data Fig. 6 | Filter function of the 1.93-THz bandpass and raw spectra from the THG experiment. a, Amplitude transmission function of a single 1.93-THz bandpass filter. **b**, As-measured spectral amplitude in arbitrary units (a.u.), as determined from the bare SiO_2 substrate (black) and from the graphene sample (red). The incident THz peak field of the fundamental at 0.68 THz was 61 kV cm^{-1}.

© 2018 Springer Nature Limited. All rights reserved.



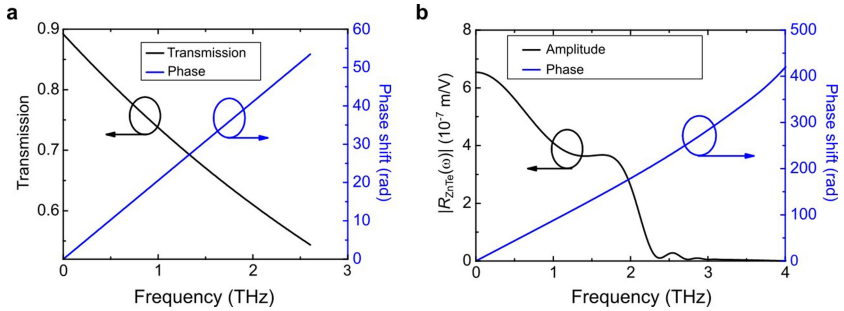
Extended Data Fig. 7 | Raw spectra from the FHG experiment. These as-measured spectra show the spectral amplitude as determined from a measurement with the bare SiO_2 substrate as a reference and a measurement of the graphene sample. The incident THz peak field in the fundamental at 0.37 THz was 40 kV cm^{-1} when using two filters in the incident beam. Insignificant transmission at the fundamental frequency and no spurious background at the FHG frequency band is observed in the reference field measurement.

© 2018 Springer Nature Limited. All rights reserved.



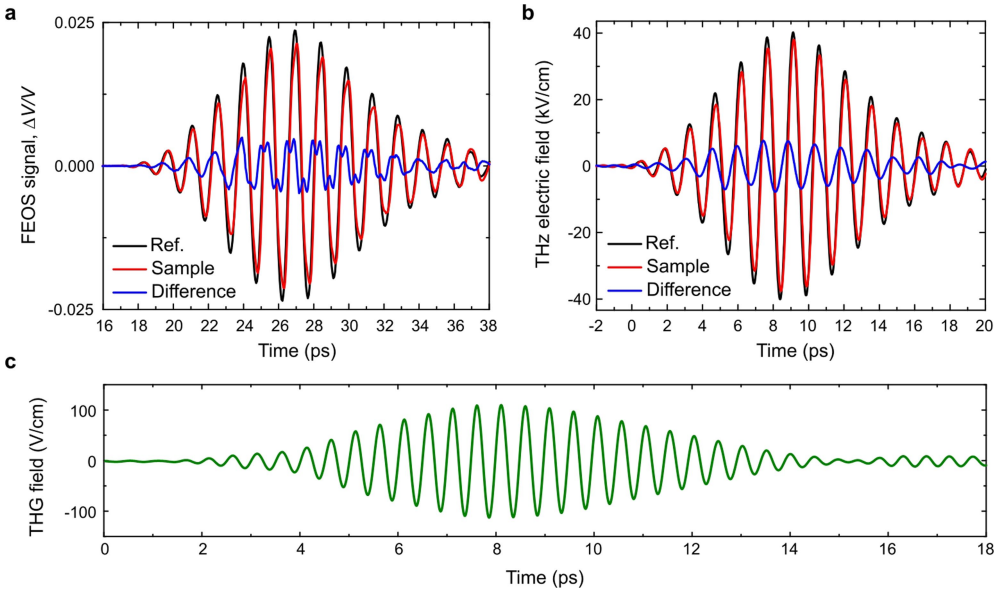
Extended Data Fig. 8 | Frequency-dependent phase difference induced by the 1.93-THz bandpass filter.

© 2018 Springer Nature Limited. All rights reserved.



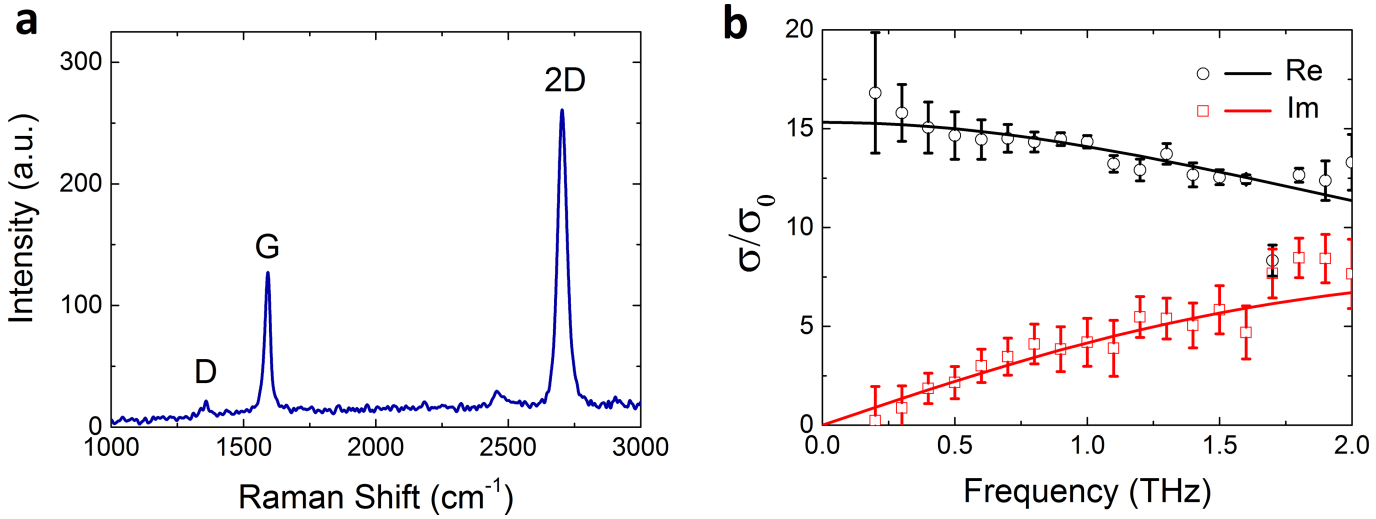
Extended Data Fig. 9 | The frequency-dependent response function. **a**, The bare substrate described by the amplitude transmission (black line) and the substrate-induced phase shift (blue line). **b**, A simulated acceptance function of the 1.9-mm-thick ZnTe detection crystal; amplitude (black curve) and phase shift (blue curve). Arrows indicate relevant axis.

© 2018 Springer Nature Limited. All rights reserved.



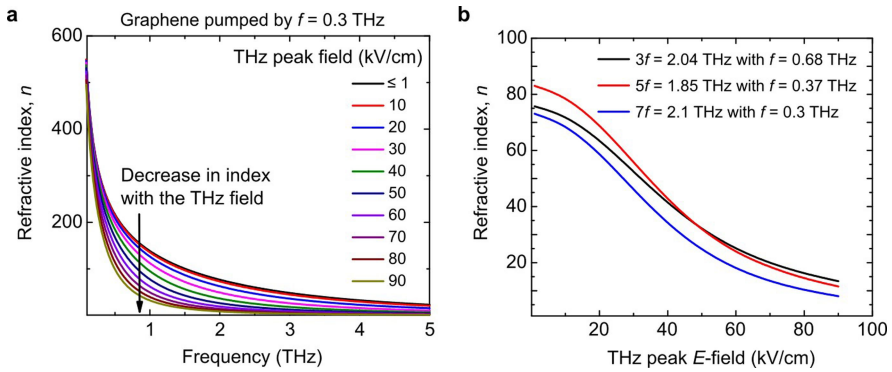
Extended Data Fig. 10 | Reconstruction of the harmonic fields from the measured FEOS signals. This is an example of THG measurement with $f = 0.68$ THz $\rightarrow 3f = 2.04$ THz. **a**, Measured FEOS signals (dimensionless). **b**, The corresponding fields transmitted through the incidence interface of the sample after deconvoluting the response functions of all the elements after the graphene film, including the 1.9-mm-thick ZnTe detection crystal, the 1.93-THz filter and the fused silica substrate from the FEOS signals in **a**. Black pulse is for the bare substrate, red for the graphene sample and blue for the difference. **c**, The pure THG field extracted from the blue field pulse in **b**.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 11 | Characterization of the graphene sample. **a**, Raman spectrum of the graphene sample. **b**, Linear conductivity, real and imaginary, of the graphene film normalized to the universal conductivity $\sigma_0 = e^2/(4\hbar)$. The symbols represent the experimental data; the solid lines represent the Drude fit with a Fermi level energy $E_F = 170$ meV (corresponding to a doping concentration $N_c = 2.1 \times 10^{12}$ cm^{-2}) and a scattering time $\tau_0 = 47$ fs as fitting parameters. The error bars are the standard deviation in the measurements.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 12 | The nonlinear (THz-field-dependent) refractive index of the graphene film. a, The THz refractive index of the graphene film as a function of frequency at various peak electric fields for the THz pump at 0.3 THz, showing reduction in the refractive index with both frequency and exciting field strength. **b**, The field dependence of the nonlinear THz refractive index at the harmonics $3f = 2.04$ THz generated by $1f = 0.68$ THz pump, $5f = 1.85$ THz generated by $1f = 0.37$ THz pump, and $7f = 2.1$ THz generated by $1f = 0.3$ THz pump.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0541-0

Multi-step self-guided pathways for shape-changing metamaterials

Corentin Coulais[1,2,3]*, Alberico Sabbadini[2], Fré Vink[2] & Martin van Hecke[1,2]

Multi-step pathways—which consist of a sequence of reconfigurations of a structure—are central to the functionality of various natural and artificial systems. Such pathways execute autonomously in self-guided processes such as protein folding[1] and self-assembly[2–5], but have previously required external control to execute in macroscale mechanical systems, provided by, for example, actuators in robotics[6–9] or manual folding in origami[8,10–12]. Here we demonstrate shape-changing, macroscale mechanical metamaterials that undergo self-guided, multi-step reconfiguration in response to global uniform compression. We avoid the need for external control by using metamaterials that are made purely of passive components. The design of the metamaterials combines nonlinear mechanical elements with a multimodal architecture that enables a sequence of topological reconfigurations caused by the formation of internal self-contacts between the elements of the metamaterial. We realize the metamaterials by using computer-controlled water-jet cutting of flexible materials, and show that the multi-step pathway and final configuration can be controlled by rational design of the nonlinear mechanical elements. We also demonstrate that the self-contacts suppress errors in the pathway. Finally, we create hierarchical architectures to extend the number of distinct reconfiguration steps. Our work establishes general principles for designing mechanical pathways, opening up new avenues for self-folding media[11,12], pluripotent materials[9,13] and pliable devices[14] in areas such as stretchable electronics and soft robotics[15].

Mechanical metamaterials are structured forms of matter that can be designed to exhibit a wide range of anomalous properties[16], including negative responses[17], topological polarization[18], non-reciprocity[19] and shape morphing[9,20]. These deviations from ordinary elasticity arise from soft deformation modes, which are encoded in the internal architectures of mechanical metamaterials[16,18]. Most metamaterials feature a single soft mode, which limits their deformations to single-step pathways[20–23]. Multi-step deformations, and other advanced functionalities, require designs that have a high-dimensional deformation space spanned by multiple soft modes[9,13,24–27]. However, when actuated, competition between these soft modes leads to frustration[11,12] and spatial decay of functionality[23]. Hence, multi-modal mechanical metamaterials have so far either been actuated without explicit control

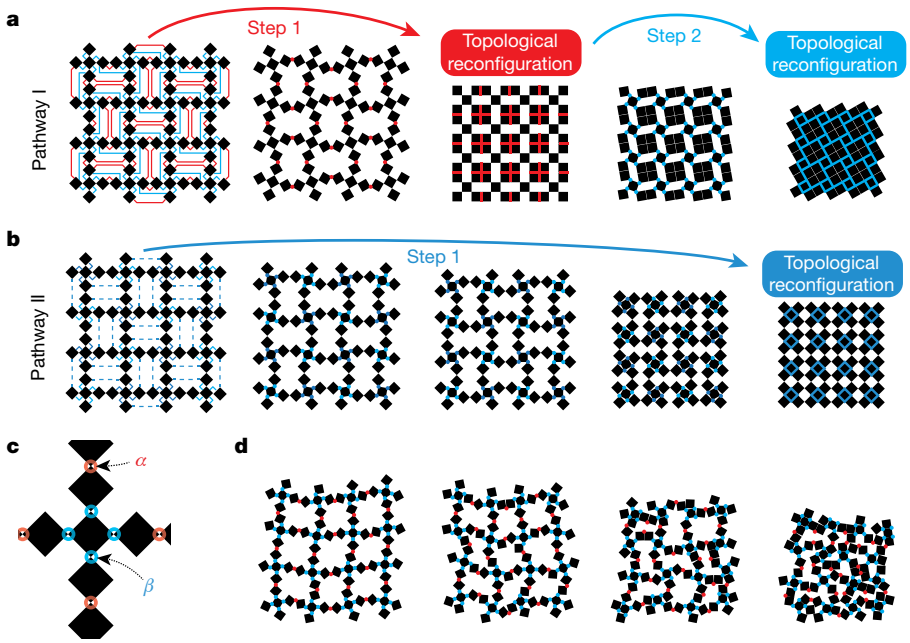


Fig. 1 | Self-guided pathways in shape-changing metamaterials. a, Two-step pathway I. In the left-most panel, the red and blue lines represent the network of self-contacts that are formed during compression. In the first step, the α links (indicated by red dots) fold and the squares connected by red lines form self-contacts (indicated in red), leading to the first topological reconfiguration; in the second step, the β links (blue dots) fold and the squares connected by blue lines form additional self-contacts (indicated in blue), leading to the second topological reconfiguration. **b**, Pathway II is a single-step pathway, in which the α links remain straight and a topological reconfiguration (blue lines) occurs after the β links are maximally folded. **c**, Close-up of the unit cell; red and blue circles indicate the α and β links and the angles α and β are defined as the deviation from the un-deformed state. **d**, Example of a disordered pathway.

[1]AMOLF, Amsterdam, The Netherlands. [2]Huygens-Kamerlingh Onnes Lab, Universiteit Leiden, Leiden, The Netherlands. [3]Institute of Physics, Universiteit van Amsterdam, Amsterdam, The Netherlands. *e-mail: coulais@uva.nl

© 2018 Springer Nature Limited. All rights reserved.

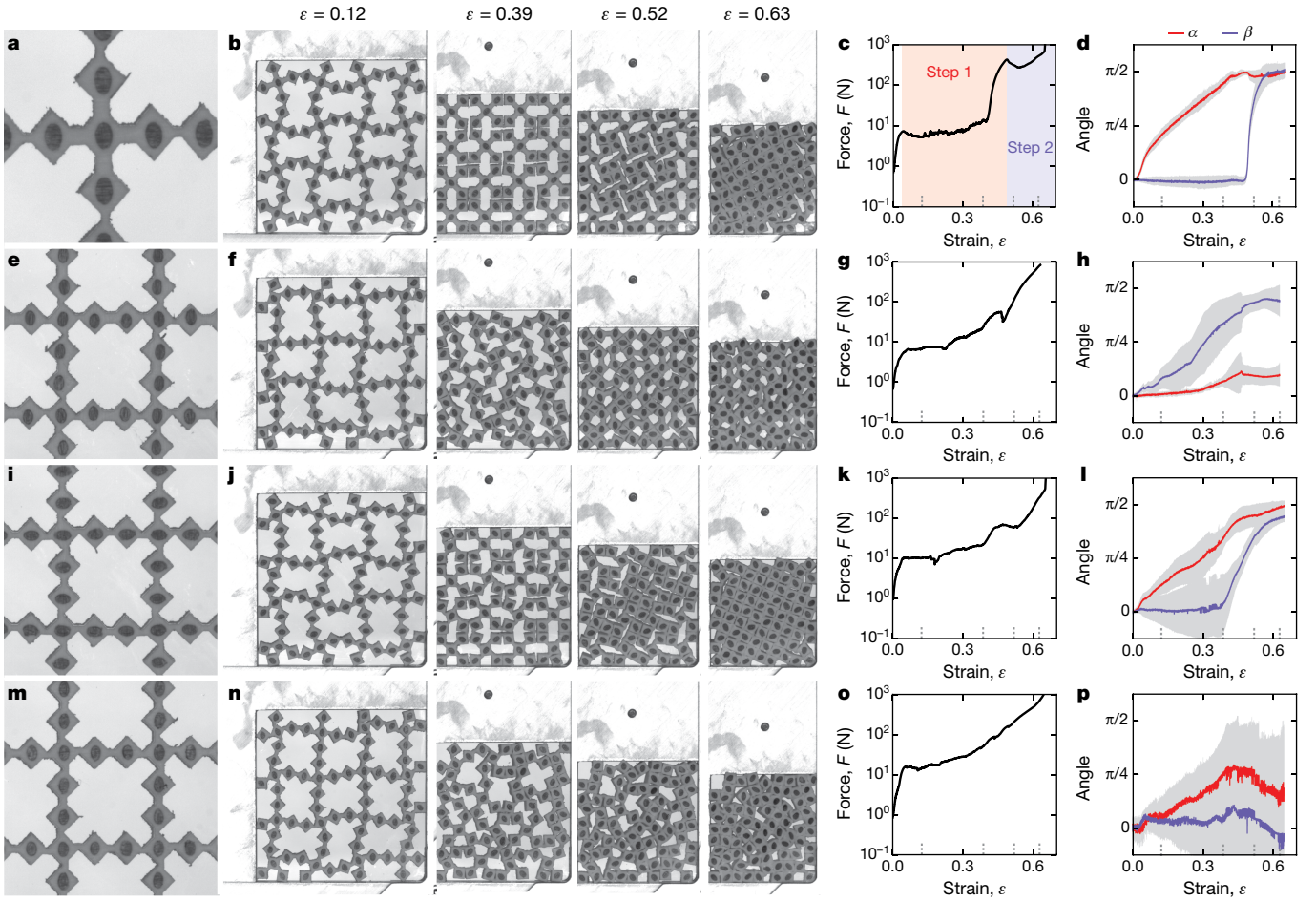


Fig. 2 | Pathway control by rational hinge design. a–d, Two-step pathway I; **e–h**, one-step pathway II; **i–l**, smoothed pathway I; **m–p**, disordered pathway. **a**, **e**, **i**, **m**, Close-ups showing hinge designs. **b**, **f**, **j**, **n**, Series of snapshots of the experimentally observed mechanical pathway, at engineering strains $\varepsilon = 0.12$, 0.39 0.52 and 0.63 (indicated by the dashed tick marks in **c**, **d**, **g**, **h**, **k**, **l**, **o** and **p**). **c**, **g**, **k**, **o**, Compressive force F versus engineering strain ε. In **c**, we indicate the two steps of the pathway. **d**, **h**, **l**, **p**, Bending of the links (α or β) versus strain ε in the central region of the sample (see Methods). See also Supplementary Video 1. The solid curves indicate the mean value of α (red) or β (blue); the grey shading indicates one standard deviation.

over their reconfiguration pathways[13,25] or required the use of multiple actuators, one for each degree of freedom[9,27]. By contrast, processes such as self-assembly[5] and protein folding[1] exhibit robust multi-step pathways without the need for external control. Here we demonstrate the design and creation of mechanical metamaterials that translate global forcing into a self-guided multi-step pathway of reconfigurations.

We consider two-dimensional metamaterials consisting of diluted lattices of freely hinged squares. We design a unit cell that consists of a cross-shaped pattern of five squares, leading to a multi-modal metamaterial of $n \times n \times 5$ squares, which allows for the formation of self-contacts between initially separated elements and has $3n^2 + 4n - 3$ zero-energy modes (Fig. 1, Methods). This structural design admits a two-step deformation pathway (which we refer to as pathway I; see Fig. 1), characterized by the deformation angles α and β (Fig. 1a, c). In the first step of pathway I, the α links fold whereas the β links remain fixed until self-contacts, which change the topology of the material, are formed. This topological reconfiguration spawns a daughter structure. In the second step, the α links remain fixed and the β links fold until the system is fully compacted (Fig. 1a). This is one possible deformation pathway, but our structure admits many other pathways, some of which are highly structured (such as pathway II; see Fig. 1b, Methods) and most of which are disordered (Fig. 1d).

To obtain a specific deformation path in response to global compression, we augment this structural design with suitably tailored mechanical interactions by connecting the squares with beam-like links of varying thickness t_α and t_β (Methods). This replaces the zero-energy modes with soft modes, the energetics of which are controlled by the links. These beams also introduce a well-defined critical compressive load proportional to t^3, above which the nonlinear buckling instability causes spontaneous bending of the corresponding link[28]. Intuitively, pathway I requires that $t_\alpha < t_\beta$, whereas pathway II requires $t_\beta < t_\alpha$ (Fig. 1). The mode structures in the corresponding asymptotic limits reveal an important qualitative difference: whereas the un-deformed metamaterial has a single soft mode for $t_\alpha \ll t_\beta$, it has many $(n^2 + 6n - 3)$ soft modes for $t_\beta \ll t_\alpha$ because the numerous soft β links can deform in many ways in this case (Methods). This counting argument suggests that pathway I will be much easier to realize than pathway II.

To design the links rationally to obtain either pathway I or pathway II, we analysed the modes and instabilities of a representative volume element—a super-cell consisting of a 2×2 grid of unit cells—as function of t_α and t_β (Methods). This modal analysis substantiates our counting-based argument. First, we find that it is the ratio t_α/t_β, rather than the individual thicknesses, that is important. Moreover, for small t_α/t_β, the first buckling mode corresponds to the first step in pathway I, with other modes strongly suppressed. Finally, for large t_α/t_β, even though the first buckling mode coincides with the first step in pathway II, other modes remain in competition. From the details of this analysis we extract specific design guidelines for the rational design of the link parameters (Methods).

We illustrate our approach experimentally. We used a water-jet cutter to fabricate reconfigurable metamaterials with squares of size 4.5 mm

© 2018 Springer Nature Limited. All rights reserved.

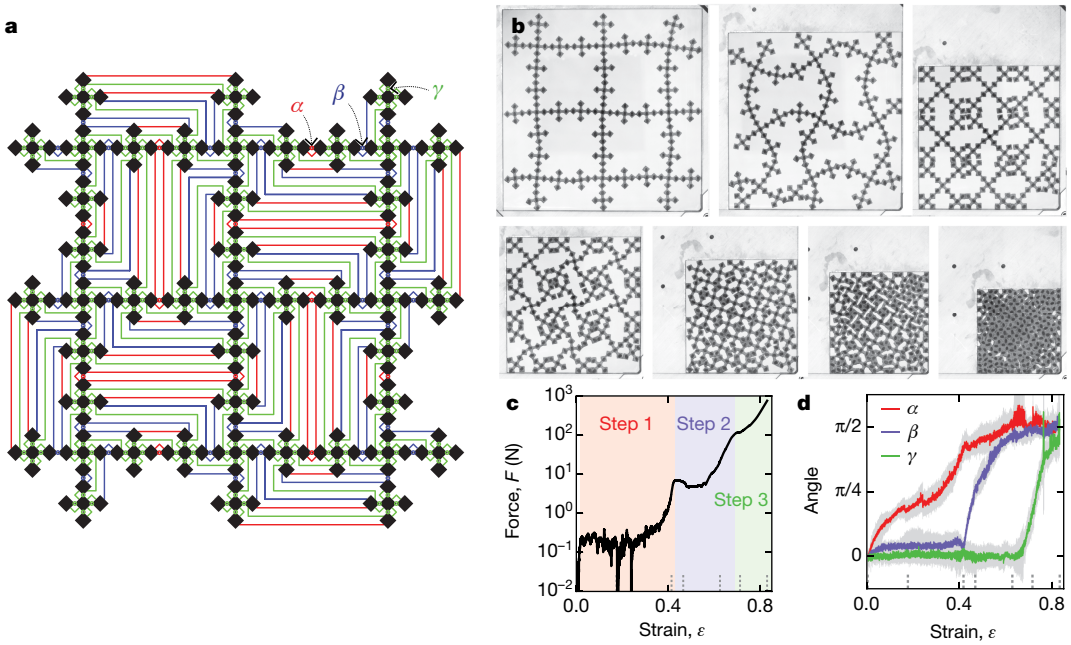


Fig. 3 | Three-step mechanical pathway. a, Geometry of a shape-changing hierarchical metamaterial. The red, blue and green coloured circles indicate the α, β and γ links, respectively. The red, blue and green coloured lines indicate the self-contacts between the squares (indicated in black) that form upon the subsequent folding of the α, β and γ links. **b**, Snapshots of the experimentally observed three-step mechanical pathway at $\varepsilon = 0$, 0.20, 0.37, 0.58, 0.62, 0.70 and 0.85 (indicated by the dashed tick marks in **c** and **d**). See also Supplementary Video 2. **c**, Compression force F versus engineering strain ε. **d**, Bending angles of the links (α, red; β, blue; γ, green) versus engineering strain ε. Lines, mean; shading, one standard deviation.

out of rubber sheets that are 10 mm thick (this prevents out-of-plane buckling). We then actuated their deformation pathways by applying a uniform equi-biaxial strain ε while tracking the compressive force, the motion of each element and the evolution of the folding angles α and β (see Methods). We focus on four different samples that share the same structural design but differ in their α and β links, which are designed to obtain one of four specific pathways (Fig. 2a, e, i, m).

We first design a metamaterial that follows the two-step pathway I. Following our design guidelines (Methods), we take $t_\alpha = 1.0$ mm and $t_\beta = 2.0$ mm (Fig. 2a). Under compression, the α links buckle at a strain of 4%, triggering the counter-rotation mode[22] and the first step of the reconfiguration pathway (Fig. 2b). Further compression does not substantially increase the force[28], preventing buckling of the stronger β links, until self-contacts are formed at a strain of 49%. The ensuing topological reconfiguration gives rise to a daughter structure with only one soft mode governed by the β links. Further compression leads to a steep rise in the compressive force, which triggers the buckling of the β links and initiates the second reconfiguration step; this second step continues until the structure is fully compressed (Fig. 2b). The compressive force clearly shows the plateaus associated with buckling and the rapid increase associated with the formation of self-contacts (Fig. 2c). This multi-step pathway is fully consistent with the theoretical pathway I, as further demonstrated by the distinct evolution of the angles α and β as functions of strain (Fig. 2d, Methods). Different geometries can be designed that have similar multi-step pathways under uniaxial compression (Methods). Hence, the combination of self-contacts, buckling and topological reconfiguration enables multi-step sequences in which buckling initiates bending, and sufficient bending—or folding—creates self-contacts and topological reconfigurations, which in turn trigger more buckling events.

Successful execution of a structured pathway requires error-correcting mechanisms that compensate for distortions caused by boundary frustration and manufacturing imperfections. Although the scatter in α and β increases considerably during each deformation step, it decreases sharply upon reaching self-contact (Fig. 2d). This reduction results from the self-alignment of the flat interfaces of contacting squares and underlies the feature of our metamaterials that the fully compressed structure is highly ordered. Hence, self-contacts are crucial for error correction, ensuring robust execution of a pathway.

Pathway II is harder to realize experimentally. Even for large $t_\alpha/t_\beta = 4$ we observe that although the deformation is qualitatively similar to pathway II it is disordered, which we attribute to competition between multiple modes (Methods). To address this issue, we demonstrate that symmetry-breaking perturbations can suppress the undesired modes and promote pathway II. We introduce another design parameter, δ_β, and laterally offset the β links by distance $\delta_\beta = t_\beta/2$ with alternating chirality (Fig. 2e, Methods). The design guideline that we extracted from our numerical analysis suggests that such symmetry-breaking offsets are sufficient to favour pathway II for any t_α/t_β greater than about 2 (Methods). We therefore created a second sample with these offsets, and $t_\alpha = 2.0$ mm $> t_\beta = 0.8$ mm. We find that under compression the slender β links bend and the thicker α links remain essentially straight, resulting in a single-step pathway to a different end state (Fig. 2e–h).

Offsets can also be used to modify and probe pathway I. By fixing $\delta_\beta = t_\beta/2$ with constant chirality (Fig. 2i, m) and gradually increasing t_α/t_β, we find experimentally that up to $t_\alpha/t_\beta \approx 1.2$ the broken symmetry causes blurring of the pathway and larger scatter in the bending angles; however, we still obtain pathway I and the fully folded state is still reached, which illustrates the robustness of pathway I against perturbations (Fig. 2i–l; $t_\alpha = 1.2$ mm, $t_\beta = 1.0$ mm). For larger t_α/t_β (more than about 1.4), we observe a disordered pathway (Fig. 2m–p; $t_\alpha = 1.4$ mm, $t_\beta = 1.0$ mm). These examples with offset links demonstrate that symmetry breaking provides an efficient strategy to steer the deformation pathway.

Finally, the hierarchical nature of our structural design can be used to create metamaterials with a three-step pathway. Our two-step design results from replacing each square in a rotating square mechanism[21] with a cross-shaped pattern of five smaller squares (Methods). To obtain a three-step sequence, we replace each square in the unit cell of the two-step design with a smaller cross-shaped subpattern (Methods). The resulting unit cell consists of 25 small squares that form a generation-two box fractal; we create a metamaterial out of nine of these unit cells, which, in the limit of perfect hinges, has 216 zero modes (Fig. 3a,

© 2018 Springer Nature Limited. All rights reserved.

Methods). We obtain a three-step pathway by sequential folding of the α, β and γ links, which leads to an intricate web of self-contacts (Fig. 3a, Methods). We tailor the three classes of links to set up this pathway by taking $t_\alpha = 1.0\ \text{mm} < t_\beta = 1.2\ \text{mm} < t_\gamma = 2.0\ \text{mm}$. Upon biaxial compression, we observe a three-step sequence of reconfigurations that leads to a fully closed, highly ordered structure (Fig. 3b, Supplementary Video 2). The compressional force and bending angles further evidence the three distinct steps of this reconfiguration pathway (Fig. 3c, d), demonstrating that our design strategy can be used to achieve complex, multiple-step pathways that are very robust.

We have shown that hierarchical structures dressed with mechanical switches based on buckling and self-contacts undergo self-guided topological reconfigurations that are robust to imperfections. Our designs are purely geometric and can therefore be applied over a range of scales; they are also passive, alleviating the need for external control. We expect our approach to enhance the potential of, for example, reconfigurable materials[9,13,29,30], soft robotics[8], origami metamaterials[10–12] and stretchable electronics[13,14].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0541-0.

Received: 14 January 2018; Accepted: 13 August 2018;
Published online 26 September 2018.

1. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
2. Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394**, 539–544 (1998).
3. Wei, B., Dai, M. & Yin, P. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* **485**, 623–626 (2012).
4. Chen, Q., Bae, S. C. & Granick, S. Directed self-assembly of a colloidal kagome lattice. *Nature* **469**, 381–384 (2011).
5. Zeravcic, Z. & Brenner, M. P. Self-replicating colloidal clusters. *Proc. Natl Acad. Sci. USA* **111**, 1748–1753 (2014).
6. Zykov, V., Mytilinaios, E., Adams, B. & Lipson, H. Robotics: self-reproducing machines. *Nature* **435**, 163–164 (2005).
7. Shepherd, R. F. et al. Multigait soft robot. *Proc. Natl Acad. Sci. USA* **108**, 20400–20403 (2011).
8. Felton, S., Tolley, M., Demaine, E., Rus, D. & Wood, R. Applied origami. a method for building self-folding machines. *Science* **345**, 644–646 (2014).
9. Overvelde, J. T., Weaver, J. C., Hoberman, C. & Bertoldi, K. Rational design of reconfigurable prismatic architected materials. *Nature* **541**, 347–352 (2017).
10. Waitukaitis, S., Menaut, R., Chen, B. G. & van Hecke, M. Origami multistability: from single vertices to metasheets. *Phys. Rev. Lett.* **114**, 055503 (2015).
11. Pinson, M. B. et al. Self-folding origami at any energy scale. *Nat. Commun.* **8**, 15477 (2017).
12. Stern, M., Pinson, M. & Murugan, A. The difficulty of folding self-folding origami. *Phys. Rev. X* **7**, 041070 (2017).
13. Cho, Y. et al. Engineering the shape and structure of materials by fractal cut. *Proc. Natl Acad. Sci. USA* **111**, 17390–17395 (2014).
14. Xu, S. et al. Materials science: assembly of micro/nanomaterials into complex, three-dimensional architectures by compressive buckling. *Science* **347**, 154–159 (2015).
15. Rogers, J., Huang, Y., Schmidt, O. G. & Gracias, D. H. Origami mems and nems. *MRS Bull.* **41**, 123–129 (2016).
16. Bertoldi, K., Vitelli, V., Christensen, J. & van Hecke, M. Flexible mechanical metamaterials. *Nat. Rev. Mater.* **2**, 17066 (2017).
17. Bertoldi, K., Reis, P. M., Willshaw, S. & Mullin, T. Negative Poisson's ratio behavior induced by an elastic instability. *Adv. Mater.* **22**, 361–366 (2010).
18. Huber, S. D. Topological mechanics. *Nat. Phys.* **12**, 621–623 (2016).
19. Coulais, C., Sounas, D. & Alú, A. Static nonreciprocity in mechanical metamaterials. *Nature* **542**, 461–464 (2017).
20. Coulais, C., Teomy, E., de Reus, K., Shokef, Y. & van Hecke, M. Combinatorial design of textured mechanical metamaterials. *Nature* **535**, 529–532 (2016).
21. Grima, J. N. & Evans, K. E. Auxetic behavior from rotating squares. *J. Mater. Sci. Lett.* **19**, 1563–1565 (2000).
22. Mullin, T., Deschanel, S., Bertoldi, K. & Boyce, M. C. Pattern transformation triggered by deformation. *Phys. Rev. Lett.* **99**, 084301 (2007).
23. Coulais, C., Kettenis, C. & van Hecke, M. A characteristic length scale causes anomalous size effects and boundary programmability in mechanical metamaterials. *Nat. Phys.* **14**, 40–44 (2018).
24. Gatt, R. et al. Hierarchical auxetic mechanical metamaterials. *Sci. Rep.* **5**, 8395 (2015).
25. Tang, Y. et al. Design of hierarchically cut hinges for highly stretchable and reconfigurable metamaterials with enhanced strength. *Adv. Mater.* **27**, 7181–7190 (2015).
26. Mousanezhad, D. et al. Hierarchical honeycomb auxetic metamaterials. *Sci. Rep.* **5**, 18306 (2015).
27. Overvelde, J. T. et al. A three-dimensional actuated origami-inspired transformable metamaterial with multiple degrees of freedom. *Nat. Commun.* **7**, 10929 (2016).
28. Coulais, C., Overvelde, J. T. B., Lubbers, L. A., Bertoldi, K. & van Hecke, M. Discontinuous buckling of wide beams and metabeams. *Phys. Rev. Lett.* **115**, 044301 (2015).
29. Blees, M. K. et al. Graphene kirigami. *Nature* **524**, 204–207 (2015).
30. Zhang, Y. et al. A mechanically driven form of kirigami as a route to 3D mesostructures in micro/nanomembranes. *Proc. Natl Acad. Sci. USA* **112**, 11757–11764 (2015).

Acknowledgements We thank J. Mesman and D. Ursem for technical support. We acknowledge Z. Zeravcic and A. Murugan for discussions. We acknowledge funding from the Netherlands Organization for Scientific Research through grants VICI No. NWO-680-47-609 (M.v.H.) and VENI NWO-680-47-445 (C.C).

Reviewer information *Nature* thanks L. Howell and L. Valdevit for their contribution to the peer review of this work.

Author contributions All authors conceived the project. C.C. and M.v.H. developed the theoretical models and C.C. performed the numerical simulations. C.C., A.S. and F.V. designed and carried out the experiments. C.C. and M.v.H. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0541-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0541-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Hierarchical design. The architecture of our metamaterials is based on an $n \times n$ rotating-square mechanism (Extended Data Fig. 1a). We see this as the first rank in a hierarchy, and for increasing rank replace each square building block with a subpattern of five smaller squares. This construction is known as a box fractal (Extended Data Fig. 1b–d). For an $n \times n$ pattern of units of rank m, the number of squares is $n^2 \times 5^{(m-1)}$. These squares are connected by $2n(n-1)$ α links that connect different unit cells. For $m \geq 2$ there are, in addition, four internal β links per unit; for $m \geq 3$ there are, in addition, 4×5 internal γ links. In general, going from rank $m-1$ to rank m yields $4n^2 \times 5^{(m-2)}$ additional internal links. For general $m \geq 2$, this yields

$$2n(n-1) + 4n^2 \sum_{i=0}^{m-2} 5^i = -2n + n^2[1 + 5^{(m-1)}]$$

connections.

Hinged tessellations. Hinged tessellations are freely hinging structures that can be folded into a fully area-filing structure. The rotating-square mechanism is an example of a hinged tessellation (Extended Data Fig. 2a). Considering pathways in our hierarchical metamaterials, we also encounter variations on the rotating-square mechanism with unequal squares (Extended Data Fig. 2b) and crosses (Extended Data Fig. 2c); both are hinged tessellations and can form area-filling structures with multiple self-contacts.

Zero modes. When the links between the squares are completely flexible, a simple Maxwell counting argument allows us to determine the number of zero modes. We start from three degrees of freedom per square (translation and rotation) and subtract two constraints for each connection, as well as three global degrees of freedom (global translation and rotation). This yields a simple expression for the number of internal degrees of freedom n_z:

$$n_z = n^2[5^{(m-1)} - 2] + 4n - 3 \quad (1)$$

For $m=2$, $n_z = 3n^2 + 4n - 3$, yielding 61 internal zero modes for the $n=4$, rank-2 structures that we show in Figs. 1 and 2; for $m=3$, $n_z = 23n^2 + 4n - 3$, yielding 216 internal zero modes for our $n=3$, rank-3 structure (Fig. 3).

To count the number of soft modes in the limit $t_\alpha \ll t_\beta$, we consider the case in which the β links are stiff and the α links are floppy, which results in a structure that is equivalent to a rotating-square mechanism with one degree of freedom[21]. To count the number of soft modes in the limit $t_\beta \ll t_\alpha$, for $m=2$ we consider the case in which the β links are floppy and the α links are stiff; these links then impose three instead of two constraints, so that an additional $2n(n-1)$ constraints must be subtracted from equation (1), yielding $n^2 + 6n - 3$ floppy modes.

Rational design of links. In the metamaterial, we dress α links and β links with elastic beams of thicknesses t_α and t_β. Such link thicknesses control the critical buckling strains and buckling modes, with the relative thickness t_α/t_β having a crucial role. Our aim is to design these link parameters so that for each step of the pathway the required mode dominates the deformations; for the specific pathways I and II, this requires tuning t_α and t_β to obtain the first buckling mode that initiates the deformation sequence under compression.

To gain insight in the role of the link thicknesses, we take the following general design approach. First, we calculate the spectrum of soft modes at zero strain as a function of t_α and t_β, and decompose the spatial structure of the lowest-energy mode onto a basis of 'motions', which correspond to desired (and undesired) pathway steps. Although a pathway consists of highly nonlinear deformations, this linear analysis still provides important insights into the role of the link parameters. Second, we perform a nonlinear analysis to determine the critical strain and spatial structure of the first buckling mode, as functions of the link parameters. Altogether, this analysis leads to the rational design of the links to generate a desired deformation pathway.

Motions. To focus on the essential physics, we define a system of four unit cells, organized in a 2×2 super-cell and placed in a uniformly compressed square box with periodic boundary conditions. In a general $n \times n$ system, periodic boundaries tie the nodes at the left (bottom) boundary to the nodes at the right (top) boundary, leading to $4n$ constraints. To remain compatible with the experimental boundary conditions, we require that the super-cell fits in a square box: the n edges at each side are required to be aligned, leading to an additional $2(n-1)$ constraints, and the horizontal and vertical dimensions are required to be equal, leading to a single additional constraint. Hence, in the limit of flexible hinges, the number of zero modes with square periodic boundary conditions is

$$n_z^{\text{periodic}} = n^2[5^{(m-1)} - 2] - 2n - 2$$

For an $m=2$, $n=2$ super-cell with square periodic boundary conditions, we thus find six zero modes. We define a vector $\boldsymbol{b}$ that describes the bending of the links (see Extended Data Fig. 3a, b) and construct an appropriate orthogonal basis $\boldsymbol{m}_{\text{A},\ldots,\text{F}}$ which we refer to as motions A–F (see Extended Data Fig. 3c, d). We choose motions A and B to coincide with the first step of pathways I (Fig. 1a) and II (Fig. 1b), and use the motions to characterize the soft modes and buckling modes.

Linear modes. We calculated the softest eigenmodes at zero strain as functions of t_α and t_β using a finite-element method (see Methods section 'Numerical simulations'). We fix $t_\alpha + t_\beta$ to be 1 mm, 2 mm or 3 mm and focus on the role of t_α/t_β, because we find that the sum $t_\alpha + t_\beta$ has a very minor role for the mode structures. We find that for small values of t_α/t_β one mode is much softer than all others, whereas for larger values of t_α/t_β several modes are in close competition (Extended Data Fig. 4a). Consistent with our counting argument, this finding can be understood by noting that when $t_\alpha \ll t_\beta$ only motion A is soft, whereas when $t_\beta \ll t_\alpha$ motions B–E are all soft, with the eigenmodes mixing these motions.

By characterizing the spatial structure of the softest mode by projecting it onto the motions A–F, we obtain the following scenario. First, there is a broad crossover regime in which multiple motions are important, with the crossover centred around $t_\alpha/t_\beta \approx 1.5$, consistent with the fact that there are more β links than α links. Second, for small values of t_α/t_β motion A dominates, whereas at large values of t_α/t_β motions B and C both have a role (Extended Data Fig. 4b). Together, this analysis suggests that for sufficiently small t_α/t_β, where the lowest mode is well isolated and corresponds closely to motion A, pathway I can be robustly obtained, whereas the situation for large t_α/t_β is more complex, with multiple modes and motions in competition.

Buckling modes. To fully understand the role of t_α/t_β, we performed a stepwise nonlinear analysis and calculated the critical strain at which the first mode becomes unstable, and similarly projected the corresponding nonlinear buckling mode onto the six motions A–F (Extended Data Fig. 4c, d). These data confirm a crossover at $t_\alpha/t_\beta \approx 1.5$. The buckling modes can be expressed almost completely as a combination of motions A and B, with other motions suppressed and essentially irrelevant.

Symmetry-broken links. Execution of pathway II is more difficult than pathway I, presumably because of the large number of competing modes and motions with deformations of the β links. However, it is possible to select pathway II, even for moderate values of the thickness ratio t_α/t_β, by introducing a symmetry-breaking lateral offset in the β links. We fix the offset magnitude of this additional design parameter to be $|\delta_\beta| = 0.5t_\beta$ (so the symmetry breaking is either on or off) and focus on offset patterns consistent with motion B. We then run full nonlinear simulations on a 2×2 super-cell with square periodic boundary conditions for a range of values of t_α/t_β while $t_\alpha + t_\beta = 3$ mm is fixed. To illustrate the general scenario, we show snapshots of the un-deformed and deformed system for $t_\alpha/t_\beta = 0.4$ and $t_\alpha/t_\beta = 2.3$ in Extended Data Fig. 5a, b. In both cases there is no instability and the deformation is smooth. At large strains we observe the deformed state corresponding to motion A for $t_\alpha/t_\beta = 0.4$, illustrating that motion A is favoured for small t_α/t_β even in the presence of offsets that promote motion B. However, for larger t_α/t_β the large strain deformation corresponds to motion B, illustrating that a combination of symmetry-breaking offsets and large t_α/t_β allows pathway II to be initiated (Extended Data Fig. 5c, d).

Design guidelines and experimental validation. We now turn our numerical observations of the crossovers between different behaviours as function of t_α/t_β, with and without offsets, into practical design guidelines, noting that there are no sharp boundaries. First, we expect that pathway I should be observed without offsets whenever t_α/t_β is less than about 1, with increasing fidelity for smaller ratios; for example, for $t_\alpha/t_\beta = 0.5$, the overlap between the buckling mode and motion A is more than 90% (Extended Data Fig. 4d). Second, to obtain pathway II, we suggest combining an offset and large values of t_α/t_β. For the specific offset magnitude used here ($|\delta_\beta| = 0.5t_\beta$) and an offset pattern of alternating chirality, concomitant with motion B, our data show that for t_α/t_β greater than about 2 motion B dominates; for example, for $t_\alpha/t_\beta = 2.3$, the overlap between the nonlinear mode and motion B is already 75% (Extended Data Fig. 5c, d). Finally, we note that link thicknesses much less than 1 mm are not experimentally feasible and link thicknesses larger than a few millimetres become comparable to the size of the square elements, providing practical constraints on our design space.

Experimental validation of pathway I. We performed experiments without offset, with $t_\alpha/t_\beta = 0.5$ (Fig. 2a–d) and $t_\alpha/t_\beta = 0.4$ (not shown), that show a clear pathway I; we also performed an experiment without offset and with $t_\alpha/t_\beta = 1.4$ that shows a disordered pathway (not shown). Earlier experiments with less accurately manufactured 3D-printed samples also showed clear pathway I deformations, so we conclude that pathway I is robust for t_α/t_β less than about 1. In addition, we performed experiments with an offset pattern of constant chirality, consistent with motion E, for $t_\alpha/t_\beta = 1$, 1.2, 1.4, 1.5 and 2.5; the first two values lead to smoothed versions of pathway I and the last three to disordered pathways (Fig. 2).

Experimental validation of pathway II. We attempted to obtain pathway II experimentally without offset and with $t_\alpha/t_\beta = 4$. Despite the fact that a large part of the sample follows motion B, these samples also display a large amount of disorder, seemingly nucleating near the boundaries, which prevents a successful execution of pathway II (Extended Data Fig. 6). We therefore performed experiments with an offset $|\delta_\beta| = 0.5t_\beta$ and an alternating chirality consistent with motion B (Fig. 2).

© 2018 Springer Nature Limited. All rights reserved.

We find that pathway II can be obtained for moderate values of $t_\alpha/t_\beta = 2$ (not shown) and $t_\alpha/t_\beta = 2.5$ (Fig. 2).

Numerical simulations. For the finite-elements simulations of a representative metamaterial element we use the commercial software Abaqus/Standard.

Model definition. We model 2×2 super-cells of our metamaterials, varying the parameters t_α, t_β and the offset δ_β (Extended Data Fig. 7), using a neo-Hookean energy density as a material model, a shear modulus $G = 2.7$ MPa, bulk modulus $K = 133$ GPa (or equivalently a Young's modulus $E = 8.0$ MPa and Poisson ratio $\nu = 0.49999$) and plane stress conditions with hybrid quadratic triangular elements (Abaqus type CPS6). We construct the mesh so that the thinnest parts of the samples are at least two elements across. As a result, the unit cells of the metamaterial have approximately 10^4 triangular elements.

Boundary conditions. To implement periodic square boundary conditions, we define constraints on the displacements of all of the nodes at the horizontal and vertical boundaries of the unit cell[31]. In addition, we ensure that the periodic boundary remains square-shaped.

Analysis. We perform three types of analysis: linear eigenmodes analysis, where we calculate the lowest eigenmodes and their eigenfrequencies (Extended Data Fig. 4a, b); nonlinear bifurcation analysis, where we perform a stepwise nonlinear stability analysis to determine the bifurcation point with a relative accuracy of 5×10^{-2} (Extended Data Fig. 4c, d); and nonlinear compression using imperfection, where to study metamaterials in which β links have an offset δ_β we compress the structure up to a strain of 6.6% (Extended Data Fig. 5).

Experimental techniques. We used 10-mm-thick cast sheets of silicone rubber (shore 80A, Silex Silicones LTD, Young's modulus $E = 8$ MPa), which are sufficiently stiff to ensure minimal affects from gravity and exhibit minimal viscous and creeping effects. We fabricated the samples using a water-jet cutter, gluing the rubber to 10-mm-thick plywood boards to ensure precise cutting. We then marked each internal square (side length of 4.5 mm) of the sample with an ellipse (2.5 mm × 4 mm) for detection purposes.

The hinges between each square consist of rectangular cuboid beams of height 10 mm (same at the sheet thickness) and lateral thickness t_α, t_β or t_γ. Irrespective of the beams, the (virtual) corners of the squares meet precisely in a point—the beams add material to this idealized design (Extended Data Fig. 7). The thicknesses t_α, t_β and t_γ range from 0.8 mm to 4.0 mm depending on the nature of the link (α, β or γ) and on the sample (see main text and Methods section 'Design guidelines and experimental validation' for specifications). In some runs, we offset the beams laterally by $\delta_\beta = t_\beta/2$, leading to triangular-shaped links. The rank-2 (rank-3) metamaterials shown in Figs. 1, 2 (Fig. 3) consist of 4×4 (3×3) unit cells, namely 80 (225) internal squares.

The samples were tested under equi-biaxial compression. To this end, we used a custom-made aluminium V-shaped press (see Extended Data Fig. 8) that was carefully positioned and aligned in a universal testing machine (Instron 3366) equipped with a 1-kN load cell, which enabled us to impose a compressive displacement with 10-μm accuracy and to record the force with 0.1-N accuracy. We used fine powder to reduce the friction at the boundaries as much as possible.

The pictures were recorded using a high-resolution complementary metal–oxide–semiconductor (CMOS) camera (3,858 pixels × 2,764 pixels, Basler acA3800-14um), equipped with a 75-mm prime lens (KOWA; LM50HC 1″) for the rank-2 metamaterial and with a 50-mm prime lens (KOWA; LM75HC 1″) for the rank-3 metamaterial, paired with custom-made LED-based front- and back-light systems. To minimize reflections from the front light, the V-press glass windows are coated.

Image tessellation techniques. The images were processed using standard image tessellation and tracking techniques to extract the positions of the squares and the bending angles of the links that connect the squares. In particular, using a semi-automatic custom-made tracking algorithm, special attention was devoted to obtain space–time trajectories of nearly all squares and bending angles (see Supplementary Videos 1 and 2). Inevitable detection errors occurred close to the edges of the compression cell and when squares came in contact, which led to flickering of the positions and orientation of the squares and therefore of the bending angles. To address this issue, events with angle fluctuations of more than 10° between two consecutive frames were filtered out. We verified that the value of the filtering threshold has only a minor effect on the data.

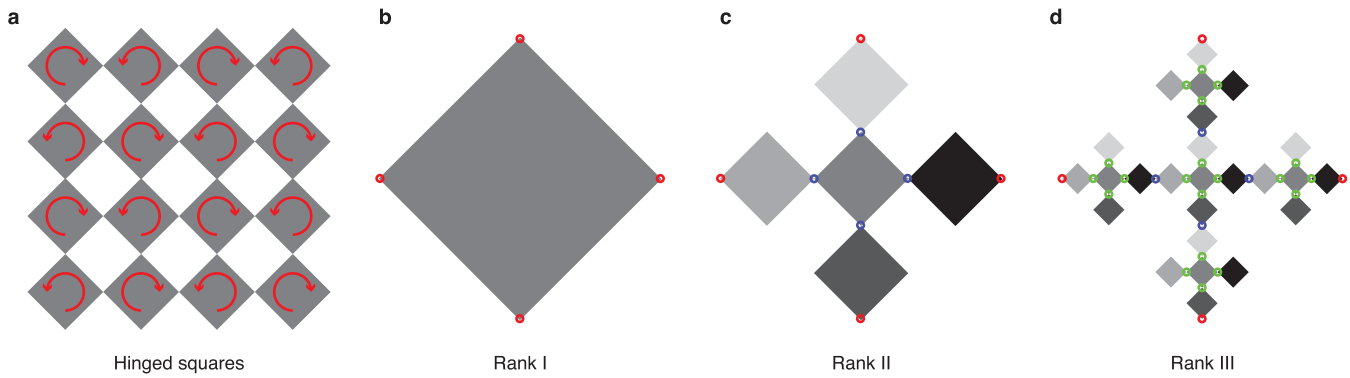
In addition, as discussed in the main text, friction and frustration at the boundary inevitably cause distortions and misfoldings, the description of which is beyond the scope of this work. Therefore, to compute the bending angles shown in Fig. 2d, h, l, p (Fig. 3), we restrict our attention to a region of interest comprising the 28 (29) most central squares for the rank-2 (rank-3) metamaterial. In Extended Data Fig. 9 we show the same data, but calculated over a much wider area, comprising 64 (165) squares. Even though the scatter is larger, the main trends and multi-step nature are still apparent in these data.

Sequential pathways with alternative topologies. To demonstrate that sequential pathways are not limited to the hierarchical structures under biaxial compression shown in the main text, we constructed two alternative geometries that exhibit a two-step pathway under uniaxial compression, using the same rubber and water-jetting technique as for the samples shown in the main text (see Extended Data Fig. 10). The basic idea is to couple two groups of links with different buckling thresholds in series, so that under compression the two types of link buckle in sequence. In one geometry, we use a diluted square lattice consisting of two rows of coupled columns with different link thicknesses (Extended Data Fig. 10a). Under compression, we observe that the columns with the thinnest links buckle first, then fold up until self-contacts are created, and finally trigger the buckling of the other columns with thicker links. Because this structure is soft to lateral shear, we use lateral sliding boundaries. In the second geometry, we use a variation that removes the soft shear modes, so that no lateral boundaries are required; here, a clear two-step sequence is also observed (Extended Data Fig. 10b). These alternative designs demonstrate that the combination of buckling and self-contacts can generate sequential pathways in various structures.

Data and code availability. The data shown in the figures and the data and code that support the other findings of this study are available from the corresponding author on request.

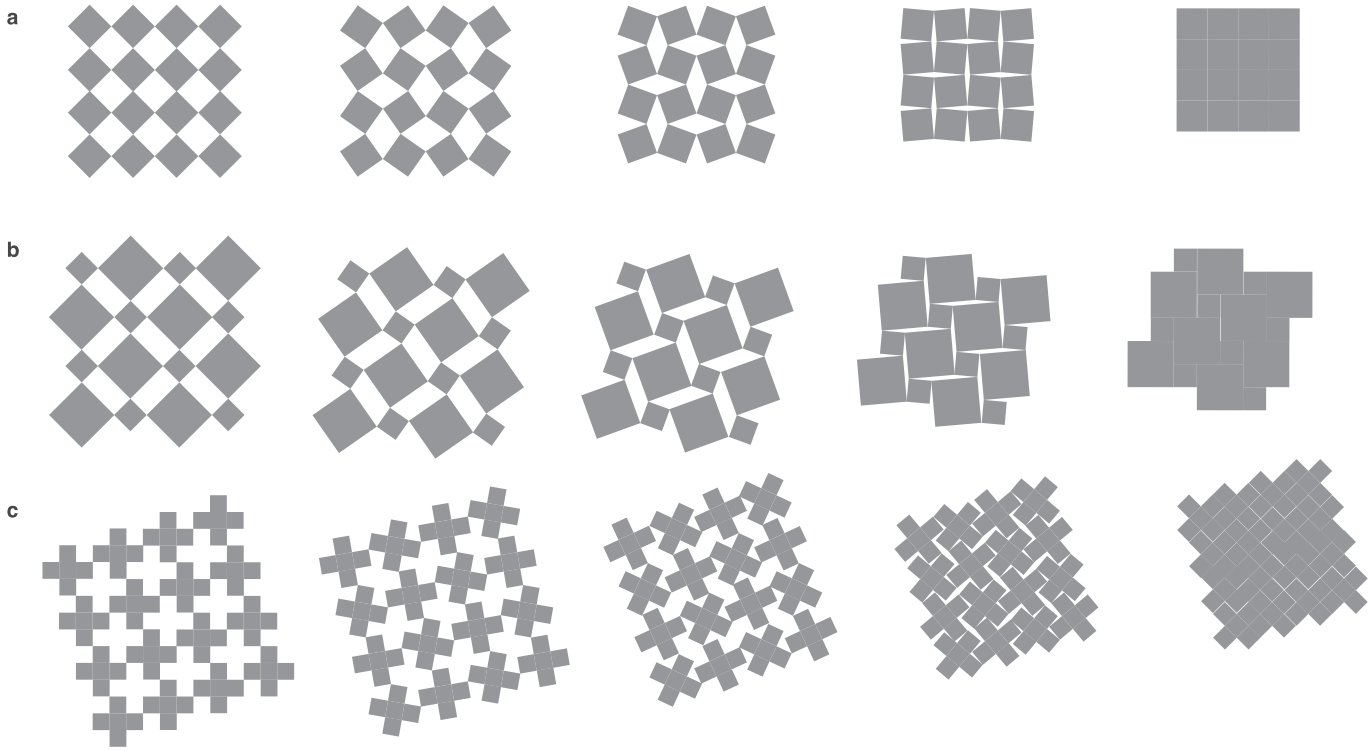
31. Coulais, C. Periodic cellular materials with nonlinear elastic homogenized stress-strain response at small strains. *Int. J. Solids Struct.* **97–98**, 226–238 (2016).

© 2018 Springer Nature Limited. All rights reserved.



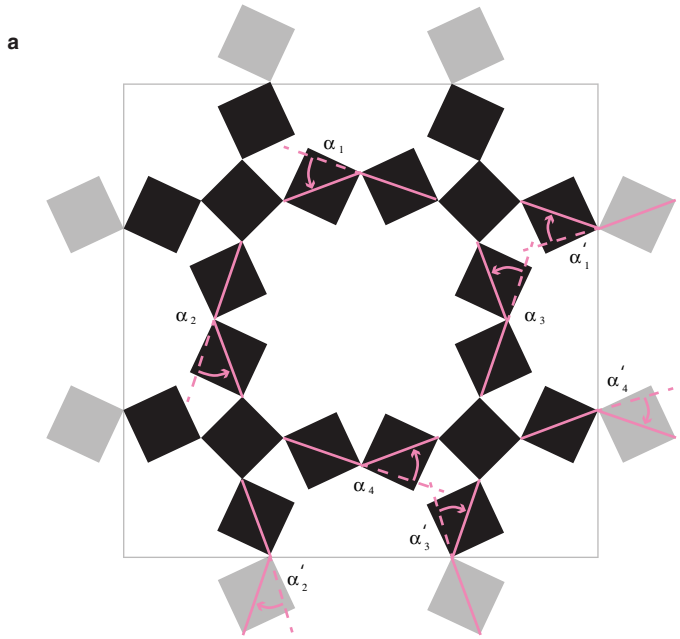
Extended Data Fig. 1 | Multimodal hierarchical mechanical metamaterials. a, 4×4 rotating-square mechanism, which has a zero mode corresponding to counter-rotation of each square unit as indicated. **b–d**, Hierarchical construction of a box fractal, where in each generation a square is replaced by a cross-like pattern of five smaller squares. Red, blue and green links correspond to α links that connect different units, internal β links that occur for rank $m \geq 2$ and γ links that occur for rank $m \geq 3$, respectively.

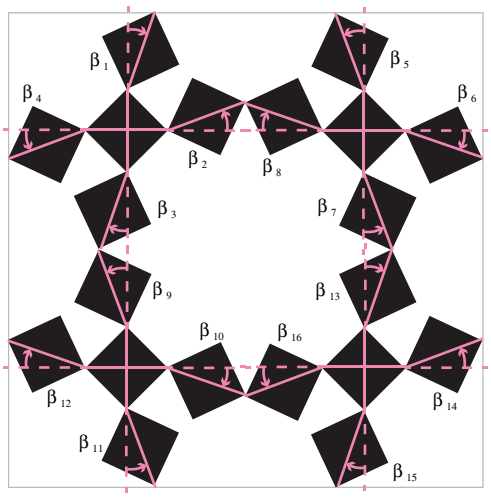
© 2018 Springer Nature Limited. All rights reserved.



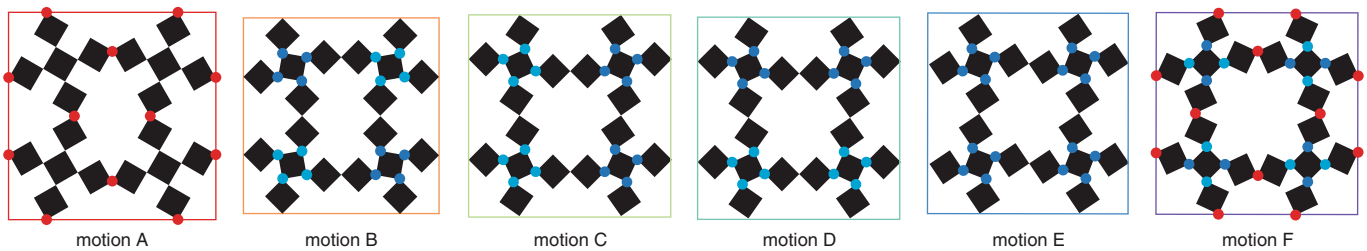
Extended Data Fig. 2 | Hinged tessellations. a, Free motion of the rotating-square mechanism. **b**, Free motion of rotating-square mechanism with unequal squares. **c**, Free motion of linked crosses.

© 2018 Springer Nature Limited. All rights reserved.





c

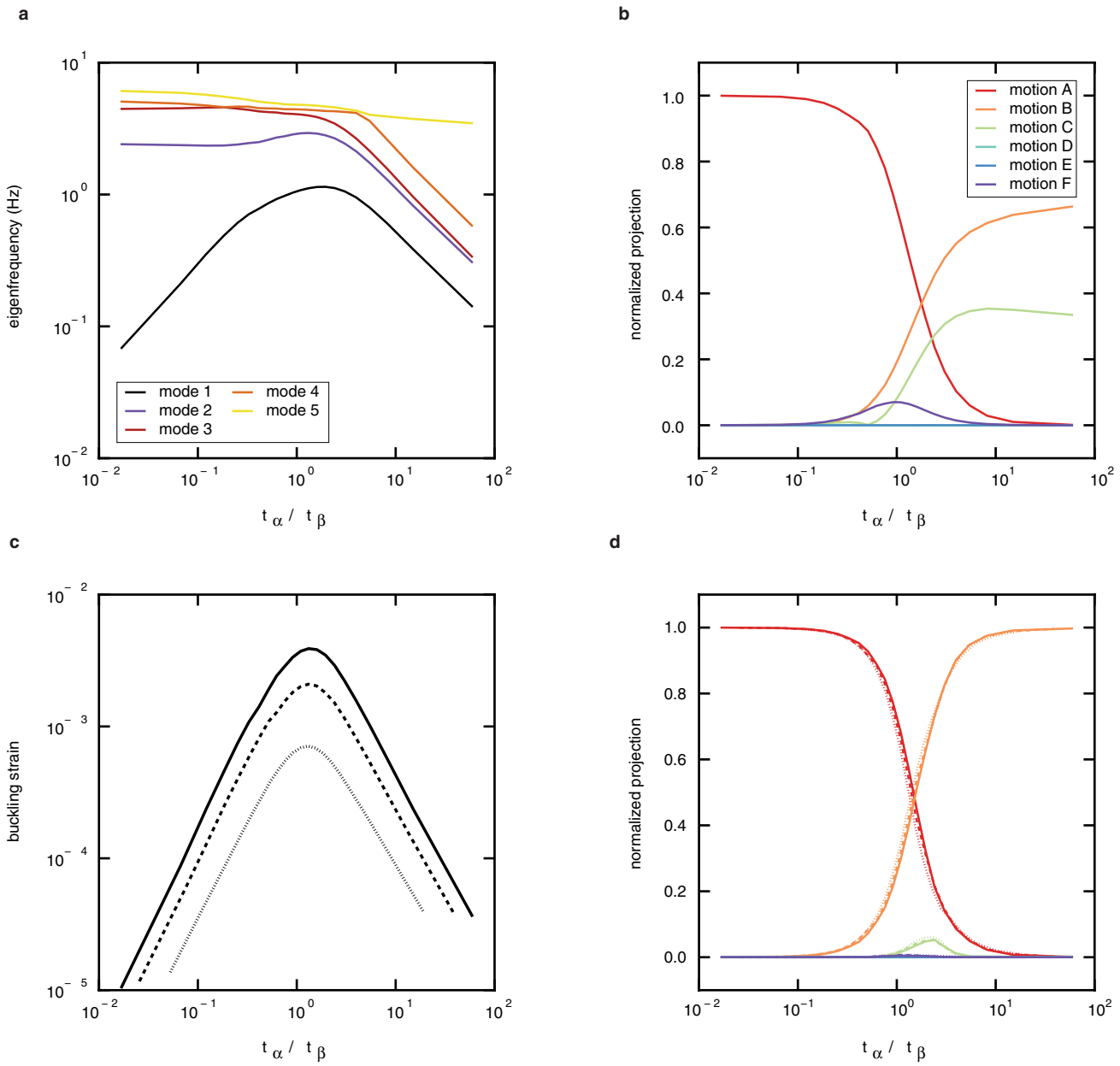


d

vectors	α_1	α_2	α_3	α_4	α'_1	α'_2	α'_3	α'_4	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
$\mathbf{m}_A$	1	−1	−1	1	−1	1	1	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathbf{m}_B$	0	0	0	0	0	0	0	0	−1	−1	−1	−1	1	1	1	1	1	1	1	1	−1	−1	−1	−1
$\mathbf{m}_C$	0	0	0	0	0	0	0	0	1	1	1	1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1
$\mathbf{m}_D$	0	0	0	0	0	0	0	0	−1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	1	1
$\mathbf{m}_E$	0	0	0	0	0	0	0	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
$\mathbf{m}_F$	2	2	2	2	−2	−2	−2	−2	−1	1	−1	1	1	−1	1	−1	1	−1	1	−1	−1	1	−1	1

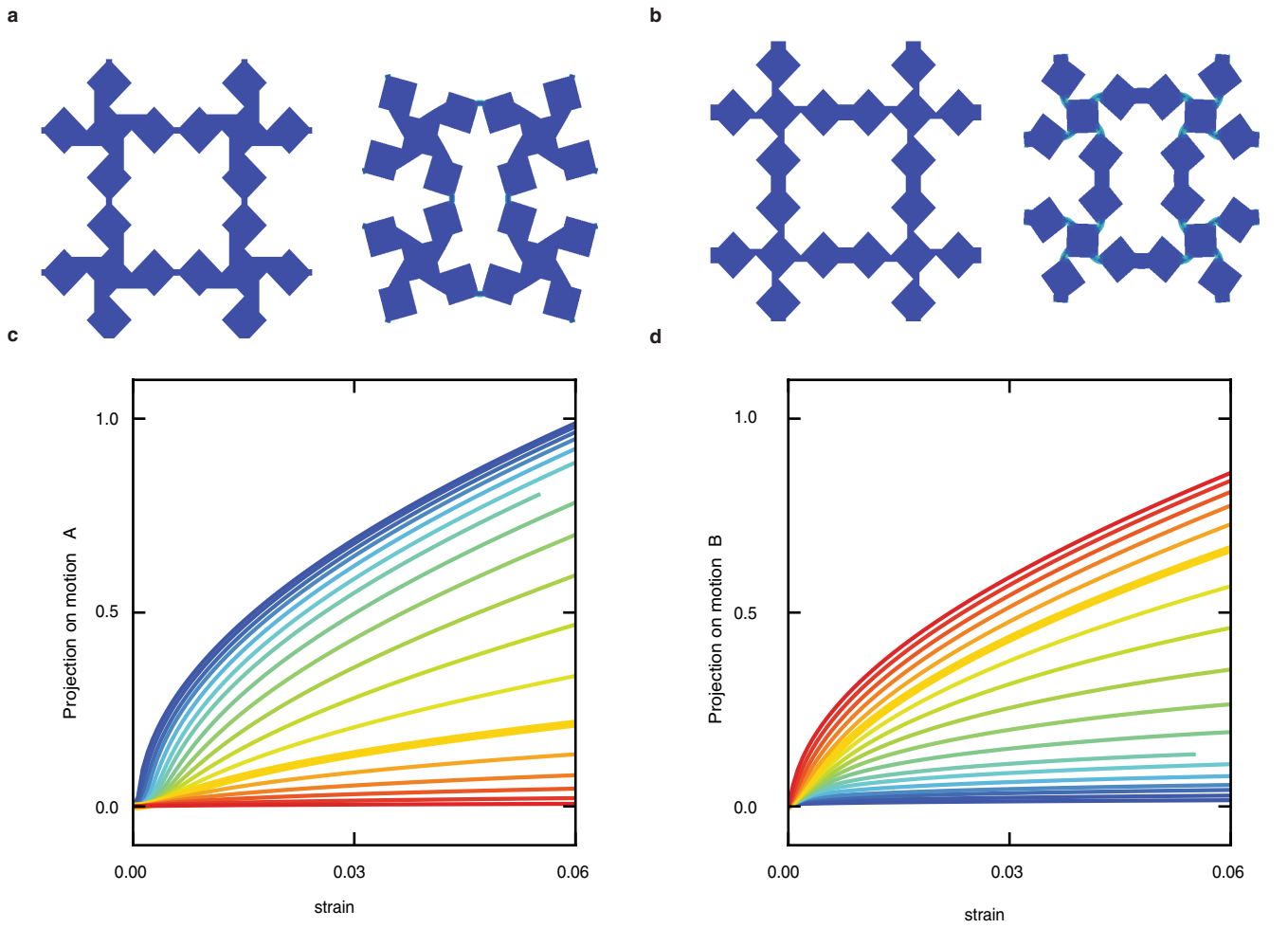
Extended Data Fig. 3 | Kinematics and orthogonal basis of zero-energy motions of a 2 × 2 super-cell with square periodic boundary conditions. **a**, Definition of the bending angles $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha'_1, \alpha'_2, \alpha'_3, \alpha'_4\}$. The grey squares depict the periodic boundary conditions. **b**, Definition of the angles $\{\beta_1, \beta_2, \ldots, \beta_{16}\}$. **c**, **d**, The motions A–F form an orthogonal basis of deformations. **c**, Schematic representation. Red denotes hinging of the α links. Light (dark) blue denotes clockwise (anticlockwise) hinging of the β links. **d**, Vector representation. The row of the table correspond to the vectors $\boldsymbol{m}_{A,\ldots,F} = \{\alpha_1, \ldots, \alpha_4, \alpha'_1, \ldots, \alpha'_4, \beta_1, \ldots, \beta_{16}\}$ that form the basis of all possible motions A–F.

© 2018 Springer Nature Limited. All rights reserved.



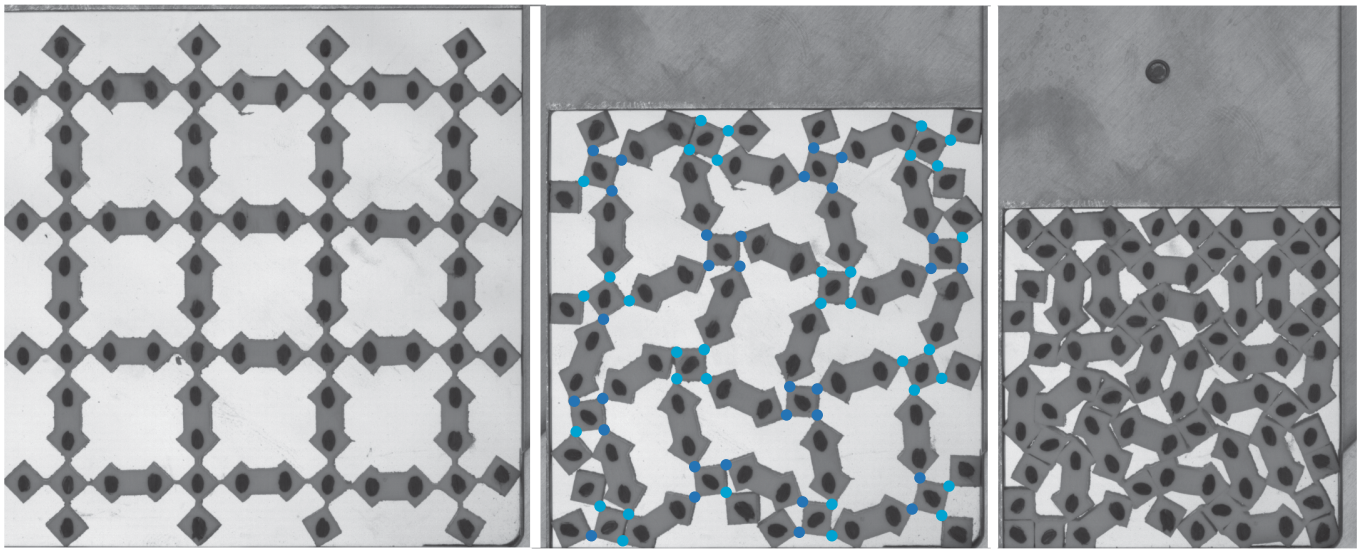
Extended Data Fig. 4 | Mode analysis for a 2 × 2 super-cell with square periodic boundary conditions. Unless noted otherwise, all data are for $t_\alpha + t_\beta = 3$ mm; data for other normalizations look extremely similar. **a**, Eigenfrequencies of the first linear eigenmodes as functions of t_α/t_β. **b**, Normalized projection of the lowest eigenmode onto motions A–F. To calculate the normalized projections, we compute the inner products between the bending-angle vectors $\boldsymbol{b}$ and the vectors $\boldsymbol{m}_{A,\ldots,F}$ defined in Extended Data Fig. 3 and normalize them by the sum of all inner products. **c**, Critical buckling strain as a function of t_α/t_β, for $t_\alpha + t_\beta = 1$ mm, 2 mm and 3 mm (dotted, dashed and solid lines, respectively). The value of the critical strain depends on $t_\alpha + t_\beta$, but the crossover does not. **d**, Normalized projection of the buckling mode onto motions A–F, for $t_\alpha + t_\beta = 1$ mm, 2 mm and 3 mm (dotted, dashed and solid lines, respectively). We calculate the normalized projections as in **c**.

© 2018 Springer Nature Limited. All rights reserved.



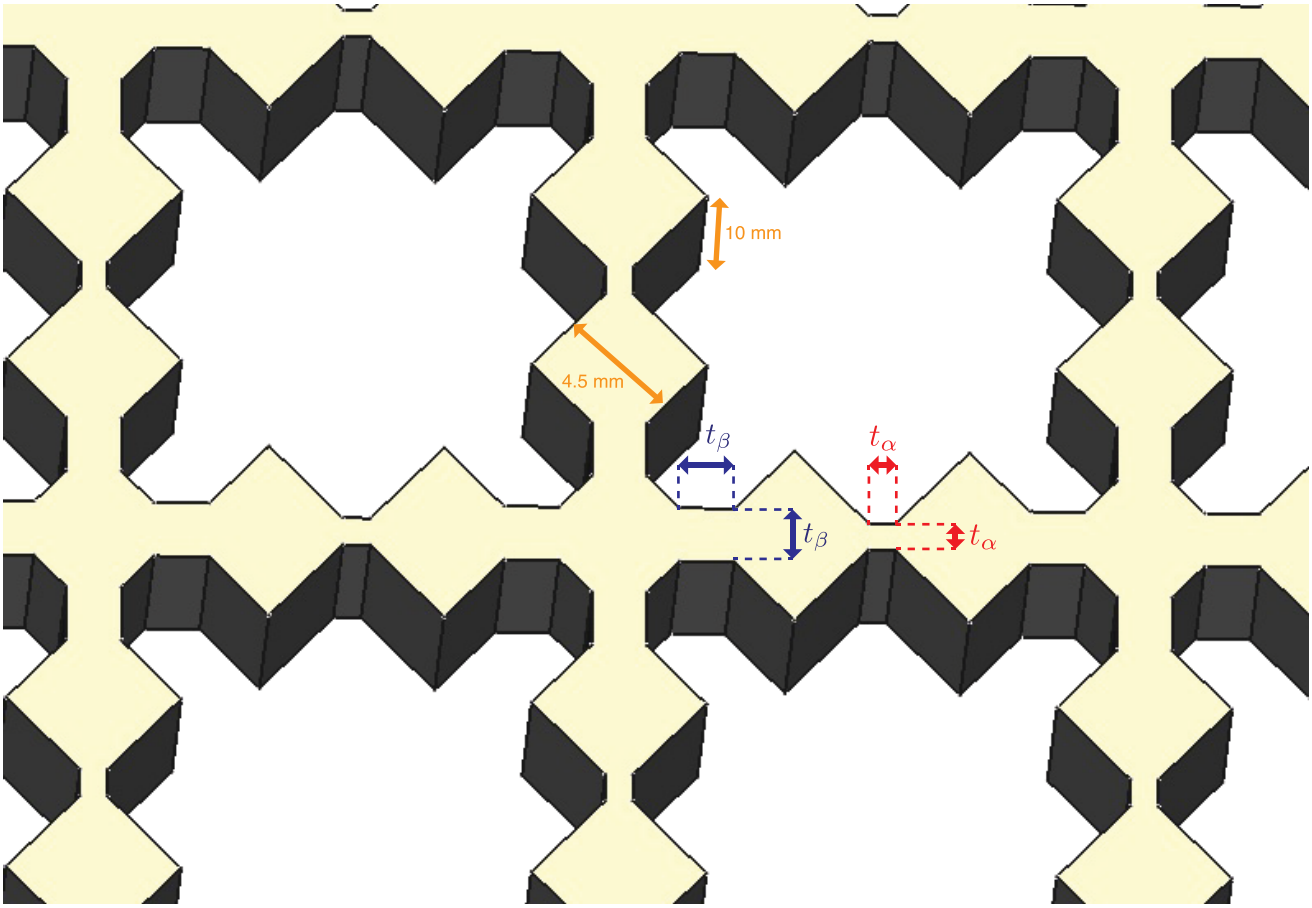
Extended Data Fig. 5 | Nonlinear analysis with symmetry-broken links on a 2 × 2 super-cell with square periodic boundary conditions. **a**, **b**, Snapshots of the super-cell in the un-deformed (left) and deformed (strain of 6.6%; right) states, for $t_\alpha = 0.9$ and $t_\beta = 2.1$ (**a**) and $t_\alpha = 2.1$ and $t_\beta = 0.9$ (**b**). In both cases, the offset of the β links is $0.5t_\beta$. **c**, **d**, Projection of the deformed states $\boldsymbol{b} \cdot \boldsymbol{m}_A$ (**c**) and $\boldsymbol{b} \cdot \boldsymbol{m}_B$ (**d**) versus strain ε for t_α/t_β ranging from 0.2 (blue) to 15 (red) and $t_\alpha + t_\beta = 3$ mm. The case $t_\alpha/t_\beta = 2.3$, which is close to the experimental value ($t_\alpha/t_\beta = 2.5$, Fig. 2e–h) is highlighted by a thick yellow line.

© 2018 Springer Nature Limited. All rights reserved.



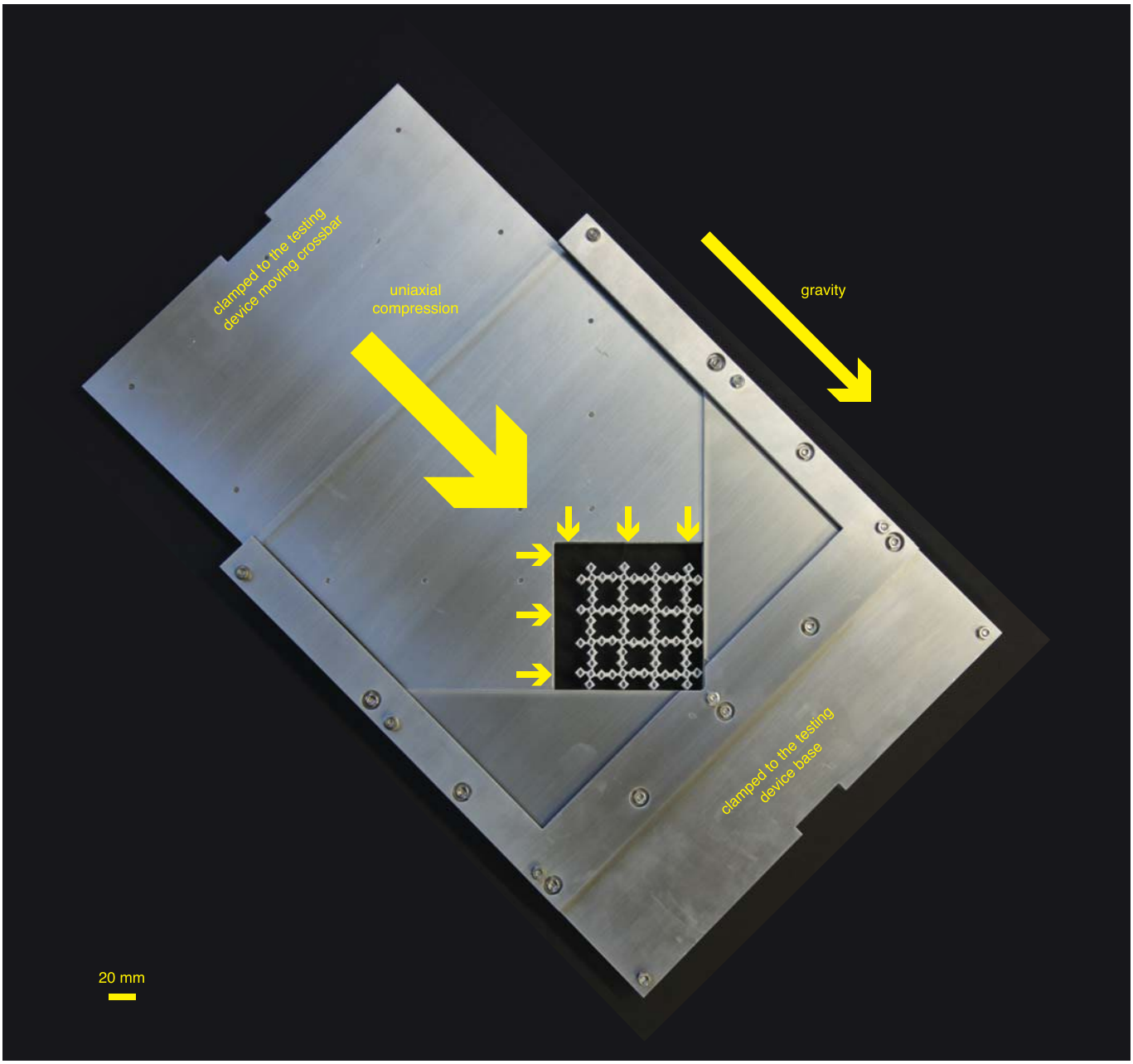
Extended Data Fig. 6 | Metamaterial under compression with $t_\alpha = 4$ mm and $t_\beta = 1$ mm. Stills of the compression experiment are shown for strains of $\varepsilon = 0$ (left), 0.24 (middle) and 0.52 (right). In the middle panel, the light (dark) blue dots indicate clockwise (anticlockwise) hinging of the β links, similar to motion B that initiates pathway II, but with substantial disorder near the boundaries that penetrate into the bulk.

© 2018 Springer Nature Limited. All rights reserved.



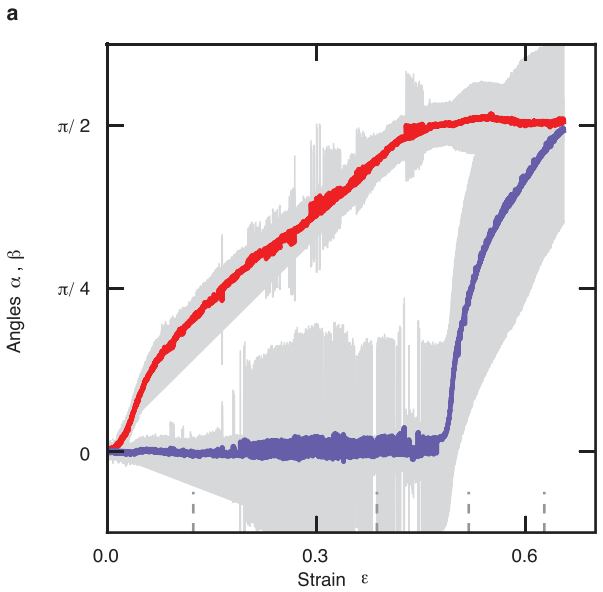
Extended Data Fig. 7 | Sample geometry. A close-up of the computer-assisted design of the rank-2 metamaterial is shown for the sample in Fig. 2a–d.

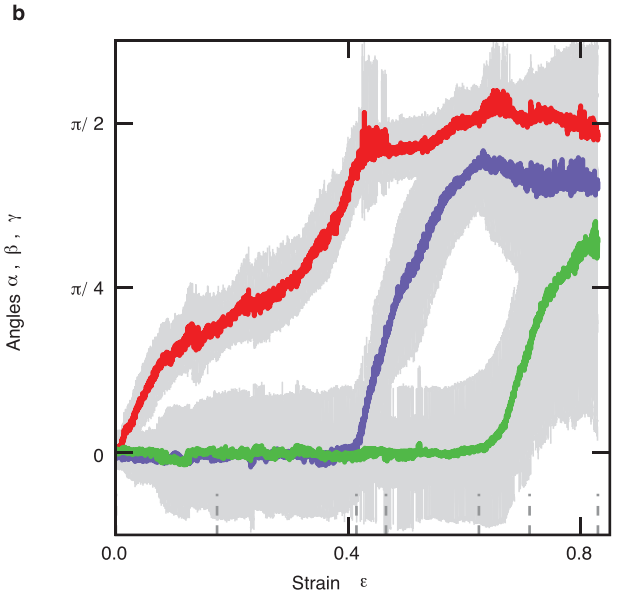
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Compression device. The custom-made compression fixture applies biaxial compression using a uniaxial testing device.

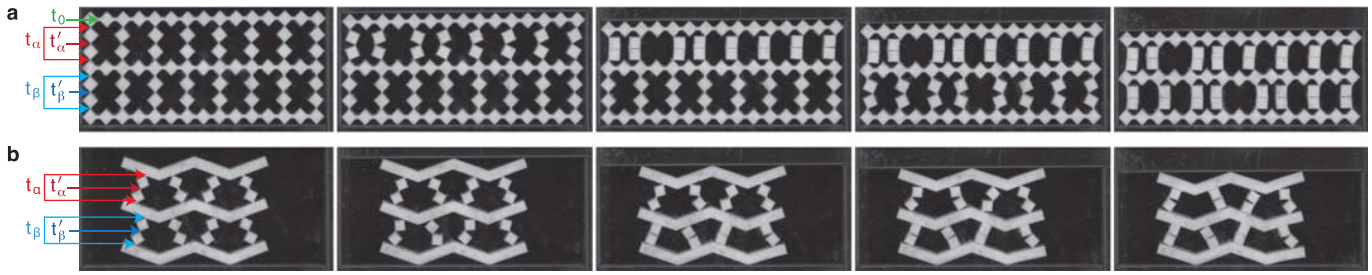
© 2018 Springer Nature Limited. All rights reserved.





Extended Data Fig. 9 | Effect of boundaries on angles. a, **b**, Bending of the links for larger regions of the sample than shown in the main text: α (red) and β (blue) versus strain ε for the central 64 squares (excluding the outermost 16 squares) (**a**); α (red), β (blue) and γ (green) versus strain ε for the central 165 squares (excluding the outermost 60 squares) (**b**). Grey shading indicates one standard deviation and the angles are based on the same runs as in Fig. 2a–d (**a**) or Fig. 3 (**b**).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Alternative topologies. a, Linked squares with edges of 4.5 mm and link thicknesses of $t_0 = 1.35$ mm, $t_\alpha = 0.45$ mm, $t'_\alpha = 0.45$ mm, $t_\beta = 0.9$ mm and $t'_\beta = 0.68$ mm, showing a two-step folding pathway under uniaxial compression. **b**, Structure of linked squares with edges of 4.5 mm and link thicknesses of $t_\alpha = 0.45$ mm, $t'_\alpha = 0.45$ mm, $t_\beta = 0.9$ mm and $t'_\beta = 0.9$ mm connected by slanted bars of thickness 4.5 mm, showing a two-step folding pathway under uniaxial compression.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0536-x

Self-powered ultra-flexible electronics via nano-grating-patterned organic photovoltaics

Sungjun Park[1,4,5], Soo Won Heo[1,5], Wonryung Lee[2,5], Daishi Inoue[1], Zhi Jiang[2,3], Kilho Yu[1], Hiroaki Jinno[1,2], Daisuke Hashizume[1], Masaki Sekino[2], Tomoyuki Yokota[2], Kenjiro Fukuda[1,3]*, Keisuke Tajima[1]* & Takao Someya[1,2,3]*

Next-generation biomedical devices[1–9] will need to be self-powered and conformable to human skin or other tissue. Such devices would enable the accurate and continuous detection of physiological signals without the need for an external power supply or bulky connecting wires. Self-powering functionality could be provided by flexible photovoltaics that can adhere to moveable and complex three-dimensional biological tissues[1–4] and skin[5–9]. Ultra-flexible organic power sources[10–13] that can be wrapped around an object have proven mechanical and thermal stability in long-term operation[13], making them potentially useful in human-compatible electronics. However, the integration of these power sources with functional electric devices including sensors has not yet been demonstrated because of their unstable output power under mechanical deformation and angular change. Also, it will be necessary to minimize high-temperature and energy-intensive processes[10,12] when fabricating an integrated power source and sensor, because such processes can damage the active material of the functional device and deform the few-micrometre-thick polymeric substrates. Here we realize self-powered ultra-flexible electronic devices that can measure biometric signals with very high signal-to-noise ratios when applied to skin or other tissue. We integrated organic electrochemical transistors used as sensors with organic photovoltaic power sources on a one-micrometre-thick ultra-flexible substrate. A high-throughput room-temperature moulding process was used to form nano-grating morphologies (with a periodicity of 760 nanometres) on the charge transporting layers. This substantially increased the efficiency of the organophotovoltaics, giving a high power-conversion efficiency that reached 10.5 per cent and resulted in a high power-per-weight value of 11.46 watts per gram. The organic electrochemical transistors exhibited a transconductance of 0.8 millisiemens and fast responsivity above one kilohertz under physiological conditions, which resulted in a maximum signal-to-noise ratio of 40.02 decibels for cardiac signal detection. Our findings offer a general platform for next-generation self-powered electronics.

The schematic of the ultra-flexible, high power-per-weight, 1D double-grating-patterned organophotovoltaic (OPV) device is displayed in Fig. 1a. For the device developed in this study, a bulk heterojunction composed of poly[4,8-bis(5-(2-ethylhexyl)thiophen-2-yl)benzo[1,2-*b*;4,5-*b*′]dithiophene-2,6-diyl-*alt*-(4-octyl-3-fluorothieno[3,4-*b*]thiophene)-2-carboxylate-2-6-diyl] (PBDTTT-OFT) polymer with [6,6]-phenyl-C71-butyric acid methyl ester ($PC_{71}BM$) (Supplementary Fig. 1) was used as the photoactive layer, and a solution-processable zinc oxide nanoparticle (ZnO NP) layer was used as the electron-transporting layer. The grating patterns for these two layers were formed by a sequential moulding process with a polydimethylsiloxane (PDMS) mould that was replicated from the surface of a blank DVD-R (having removed the cover layer; see Methods and Supplementary Figs. 2 and 3). After fabrication with parylene passivation (Supplementary Video 1), and device delamination from the supporting glass substrate, 3-μm-thick ultra-flexible OPVs with shiny outer surfaces were obtained (Fig. 1b). These were ultra-light weight (36.6 μg per unit device), with a power-conversion efficiency (PCE; 10.49%) superior to that of other flexible OPVs (Supplementary Fig. 1 and Supplementary Table 1), resulting in a high power-per-weight (11.46 W g^{-1}) that is sufficient for application as a practical mobile and portable power source. These ultra-thin devices, delaminated from the supporting glass, are referred to as free-standing OPVs.

The grating patterns were examined in both surface and cross-section (Fig. 1c, d). The period of the pattern was 760 nm and the height of the photoactive and ZnO NP layers varied from 66 to 116 nm and from 10 to 20 nm (determined from atomic force microscopy (AFM) and transmission electron microscopy (TEM) images), respectively (Supplementary Figs. 5 and 6). Figure 1e displays the current density–voltage (*J*–*V*) characteristic of the free-standing 'champion' cell, which exhibited unprecedented photovoltaic-parameter values, including a short-circuit current density (J_{SC}) of 19.17 mA cm^{-2}, an open-circuit voltage (V_{OC}) of 0.785 V and a fill factor (FF) of 0.697, resulting in a PCE of 10.49% and a high power-per-weight of 11.46 W g^{-1}. The statistical analysis of 57 devices, before delamination from the supporting glass, revealed average device parameters as follows: J_{SC}, 18.97 mA cm^{-2}; V_{OC}, 0.779 V; FF, 0.679; PCE, 10.05% (Table 1; see also Supplementary Figs. 7 and 8). The PCE enhancements of both free-standing devices (from 10.0%[13] to 10.49% for a champion cell) and rigid devices on glass substrates (from 9.29% to 9.91%, as shown in Table 1) are due to the following: (1) minimized reflection and interference of light by the decrease of the refractive index (*n*) mismatch by the patterned ZnO NP layer in the region 350–450 nm, (2) maximized light absorption between 400 and 800 nm due to scattering and the longer optical path within the photoactive layer, and (3) the extension of the absorption edge at 820 nm, attributed to the surface plasmonic effect by forming a grating on the Ag surface[14] (Supplementary Figs. 9 and 10). It is noteworthy that the double-grating patterns exhibit cumulative J_{SC} enhancement, without the severe deterioration of the other photovoltaic parameters, validating the effective combination of both the patterned layers for maximizing the optical gain.

The grating-patterned OPVs on ultra-thin parylene substrates (before delamination) exhibited a greater PCE value than that of the OPVs on glass substrates (Supplementary Fig. 11 and Supplementary Table 2), which is unusual for a conventional soft electronic device. This PCE enhancement is due to the greater absorption of transmitted light into the photoactive layer with minimized scattering because of gradual refractive index matching for the patterned ZnO NP layer ($n = 1.95$) with the insertion of parylene ($n = 1.67$) between the glass substrate ($n = 1.53$) and the indium tin oxide (ITO) electrode ($n = 1.82$)[15–17] (Supplementary Fig. 12). In addition to the high PCE, the high reproducibility of the moulding process is promising for device scalability (Supplementary Fig. 13).

Environmental stability of the 1D double-grating-patterned OPVs under conditions of mechanical and angular movement is required to

[1]Center for Emergent Matter Science (CEMS), RIKEN, Saitama, Japan. [2]Electrical and Electronic Engineering and Information Systems, The University of Tokyo, Tokyo, Japan. [3]Thin-Film Device Laboratory, RIKEN, Wako, Japan. [4]Present address: Organic Materials Laboratory, Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, Co., Suwon, South Korea. [5]These authors contributed equally: Sungjun Park, Soo Won Heo, Wonryung Lee. *e-mail: kenjiro.fukuda@riken.jp; keisuke.tajima@riken.jp; someya@ee.t.u-tokyo.ac.jp

© 2018 Springer Nature Limited. All rights reserved.

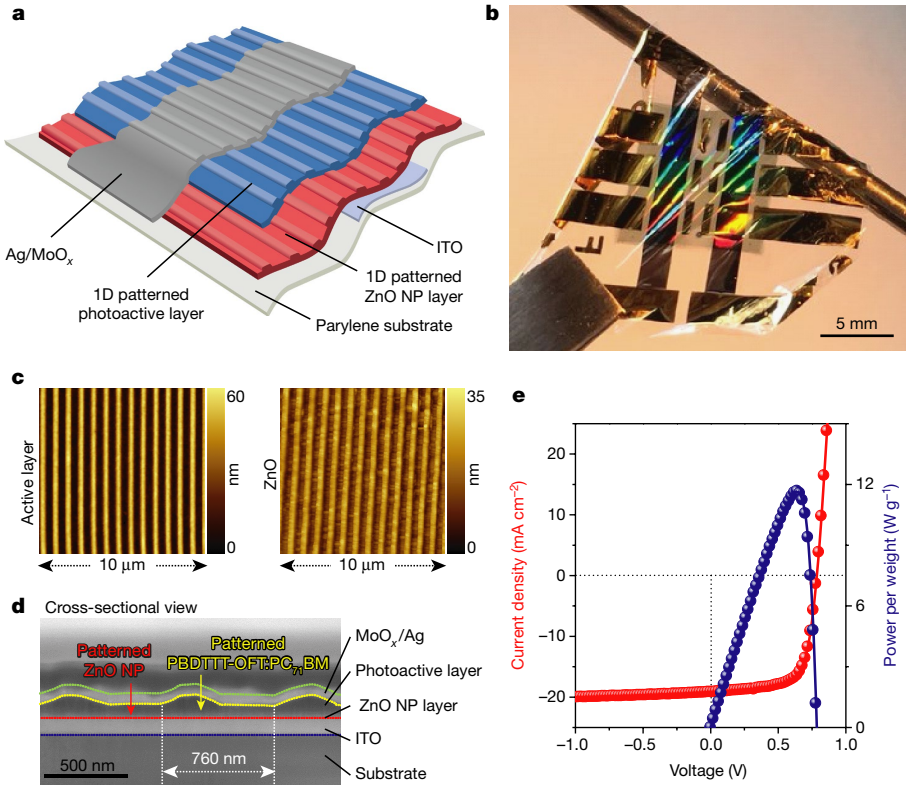


Fig. 1 | Design of the double-grating-patterned high-performance ultra-flexible OPV. a, Structure of the OPV device, see Methods for details. **b**, Photograph of the OPV device wrapped over a spatula rod and pulled by tweezers, after delamination from the supporting glass substrate. The measured weight of a single device is 36.6 μg. **c**, AFM images of the grating-patterned surface of the bulk-heterojunction layer (left) and the ZnO NP layer (right). **d**, High-resolution cross-sectional scanning electron microscopy (SEM) image of the grating-patterned structure. The nanoscale grating pattern has a grating depth of approximately 50 nm and 10 nm, for the photoactive and ZnO NP layers, respectively. **e**, Current density–voltage (*J*–*V*; red) and power-per-weight–voltage (*P*–*V*; blue) curves of the free-standing devices. The red and blue lines represent the current density measured under illumination (AM 1.5 G) and the calculated power-per-weight, respectively. Device parameters: area, 0.04 cm^2; J_{SC}, 19.17 mA cm^{-2}; V_{OC}, 0.785 V; FF, 0.697; PCE, 10.49%; power per weight, 11.46 W g^{-1}. See Methods for details.

ensure continuous supply of power to functional devices. Figure 2a displays a diagram of OPVs subjected to the compression test. After delamination from the supporting glass (Supplementary Fig. 14), the 3-μm-thick ultra-thin OPVs were laminated onto 200% pre-stretched elastomer (Supplementary Fig. 15). The devices were then compressed in the uniaxial centre direction by slowly reducing the stretching force, giving a tensile strain in the elastomer of 200%–0%, corresponding to a compression of the device. When the devices were compressed by a tensile strain of 0% in the elastomer, a high-aspect-ratio sinusoidal wrinkled surface and microscale sharp, multiple bends with a bending radius as low as 3 μm were formed (Fig. 2b and Supplementary Fig. 16). During compression (forward) and release (backward) with 10% tensile-strain steps, only the short-circuit current (I_{SC}) varied (due to the change in the effective light absorption area), while other parameters, such as V_{OC} and FF, maintained their initial values without severe deterioration (Fig. 2c–e, and Supplementary Figs. 17 and 18). This indicates that the nanoscale 1D double-grating patterns can maintain their morphologies under this sort of mechanical deformation, resulting in negligible degradation in the electrical performance. With repetitive compression cycling, it was confirmed that the PCE changed from an initial value of 9.82% to 7.33% after 900 repetitive compression cycles with 33% compression strain (Fig. 2f and Supplementary Fig. 19).

In addition to mechanical stability, the electrical characterization of the OPV under different angles of illumination is a prerequisite for mobile and portable applications. It was confirmed that patterning the structure does not affect the V_{OC} values obtained for the

Table 1 | Photovoltaic parameters of the grating-patterned OPVs

Structure	Substrate	J_{SC} (mA cm^{-2})	V_{OC} (V)	FF	PCE (%)
ZnO/active	Glass[a]	16.99 ± 0.02	0.763	71.6 ± 0.6	9.29 ± 0.14
ZnO/active$^{(P)}$	Glass[a]	17.75 ± 0.12	0.769	68.5 ± 0.3	9.34 ± 0.07
ZnO$^{(P)}$/active	Glass[a]	17.65 ± 0.06	0.759	72.4 ± 0.2	9.73 ± 0.05
ZnO$^{(P)}$/active$^{(P)}$	Glass[a]	18.56 ± 0.03	0.766	69.6 ± 0.1	9.91 ± 0.04
ZnO$^{(P)}$/active$^{(P)}$	Parylene (before delamination)[b]	18.97 ± 0.51	0.779	68.0 ± 0.01	10.05 ± 0.31
Champion cell	On parylene (before delamination)[b]	19.95	0.777	68.8	10.67
ZnO$^{(P)}$/active$^{(P)}$	Parylene (free-standing)[c]	18.50 ± 0.28	0.775	68.5 ± 0.01	9.83 ± 0.26
Champion cell	Parylene (free-standing)[c]	19.17	0.785	69.7	10.49

In the first column, $^{(P)}$indicates the DVD-moulded (patterned) layer of the photovoltaic device, 'active' indicates the photoactive layer (PBDTTT-OFT:PC$_{71}$BM). Values shown in the table are averages over the number of devices indicated in the second column, as follows.
[a]4 devices.
[b]57 devices.
[c]12 devices.

© 2018 Springer Nature Limited. All rights reserved.

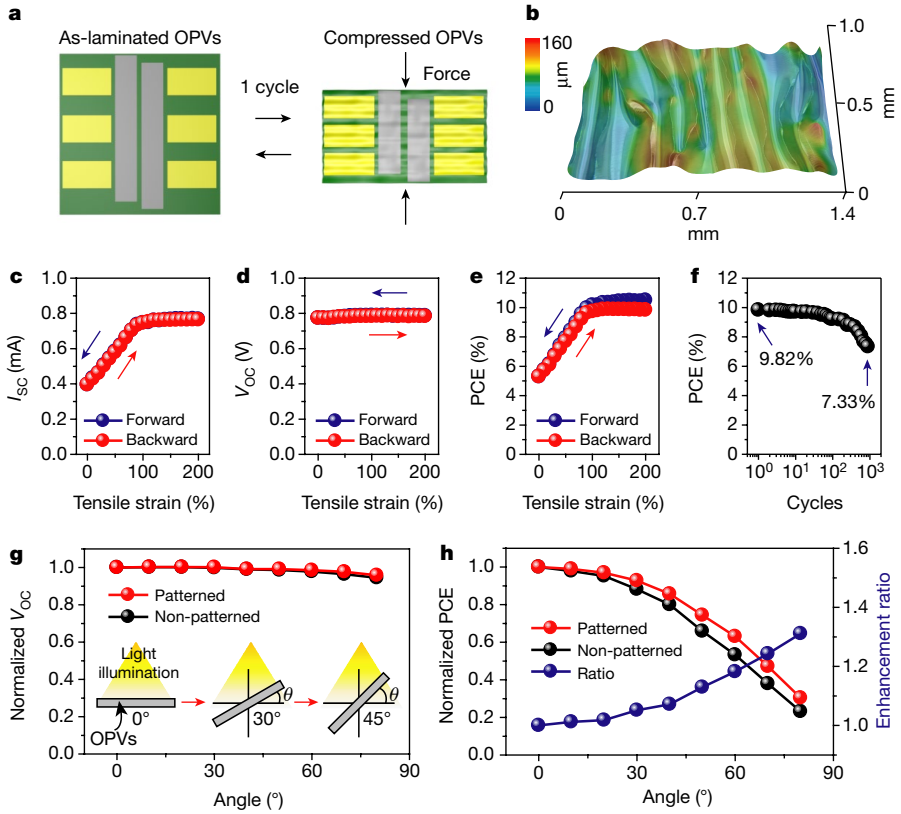


Fig. 2 | Mechanical durability of the double-grating-patterned ultra-flexible OPV. a, Schematic of the OPV, under a compression test set-up. The OPV is laminated on a stretched elastomer (left), and compressed by reducing the stretching force (right). A cycle indicates repetitive mechanical deformation by 33% mechanical compression and recovery to a flat condition. **b**, 3D microscope image of the sample undergoing 33% compressive strain. A localized sinusoidal wrinkled surface with a high aspect ratio is formed with a peak amplitude of 160 μm. **c**–**e**, Changes of the photovoltaic parameters I_{SC} (**c**), V_{OC} (**d**) and PCE (**e**) of the OPV under compression. **f**, Behaviour of the photovoltaic PCE parameter under the cyclic compression test. **g**, **h**, Normalized values of V_{OC} (**g**) and PCE (**h**) as a function of the light illumination angle. The inset of **g** is a schematic illustration of the different incident light angles in the experimental set-up.

non-patterned structure, regardless of the illumination angle (Fig. 2g), and that the values of I_{SC} and PCE obtained for the patterned structure were superior to those for the non-patterned structure for all illumination angles (Fig. 2h and Supplementary Fig. 20). The weaker angle dependence could be attributed the anti-reflection effects of the patterned ZnO[14,18], and enables sufficient power to be supplied under dynamic movement.

Taken together, these results prove that the 1D double-grating-patterned ultra-flexible OPVs are practically applicable as power sources for driving conformal functional electronic devices. We note that reduction of the angle dependence of power generation is critical for optimizing the output power of our ultra-thin OPVs, which are conformable to three-dimensional surfaces. We have been able to achieve this, for the first time to our knowledge, because nano-grating morphologies can be successfully fabricated on the surface of ZnO and polymeric active layers without applying pressure: therefore, this process is suitable for ultra-thin films.

After establishing their portability and conformal-power-source applicability, we integrated the OPVs with organic electrochemical transistors (OECTs, Supplementary Fig. 21) for demonstration as self-powered on-skin and on-tissue cardiac sensors (Fig. 3a–c and Supplementary Fig. 22). OECTs were selected because of their high flexibility (Supplementary Fig. 23), bio-compatibility, low-voltage operation and fast response speed (1 kHz)[19–23]. After device integration, we investigated the effect of light intensity on various electrical properties. Surprisingly, a high current ratio of over 10^5 was achieved between two different illumination conditions (1,000 W m^{-2} and dark; Fig. 3d), indicating that the current in the light-insensitive poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS) active channel (Supplementary Fig. 24) can be modulated by the OPVs, depending on the external light intensity (Fig. 3e and Supplementary Fig. 25). In addition, the highest transconductance value was 0.8 mS at 1,000 W m^{-2}, comparable to typical OECT devices biased with several hundred millivolts at their drain electrodes using external source meters[22]. When the light intensity was decreased to 100 W m^{-2} (approximately 10^4 lux), the transconductance value of 0.4 mS was maintained, indicating that the integrated devices can be operated as sensors under room light conditions (Fig. 3f and Supplementary Fig. 26). It is also noteworthy that the 1D double-grating-patterned OPVs exhibited higher output current than the non-patterned devices, resulting in high transconductance values for a wide light intensity range (Supplementary Fig. 27), indicating that more sensitive and efficient devices can be obtained via the grating patterning process for extracting fine biological signals. In addition, both drain current (I_D)–gate voltage (V_G) curves (Fig. 3g) and dynamic I_D output characteristics (Fig. 3h, i) were unaffected by experimental conditions (Supplementary Fig. 28), and the rising (τ_{rise}) and falling (τ_{fall}) response times were found to be approximately 0.6 ms: together, these results prove that the present devices can be applied as biological signal detectors with high-speed responses, compatible with the frequency ranges of human biological signals (approximately 200 Hz)[24].

In addition to high transconductance and high responsivity (Fig. 3), signal acquisition with minimized noise is crucial to obtaining a high signal-to-noise ratio. The standard deviation of the noise current (Fig. 4a) was found to be 0.51 μA, 0.08 μA and 0.02 μA for devices powered by a source meter, by a commercial battery and by an OPV, respectively. The integrated devices powered by light illumination exhibited much lower noise, with a lower standard deviation, than devices biased by a commercial battery, as confirmed by the statistical distribution and fast Fourier transform analysis (Supplementary Fig. 29): this good

© 2018 Springer Nature Limited. All rights reserved.

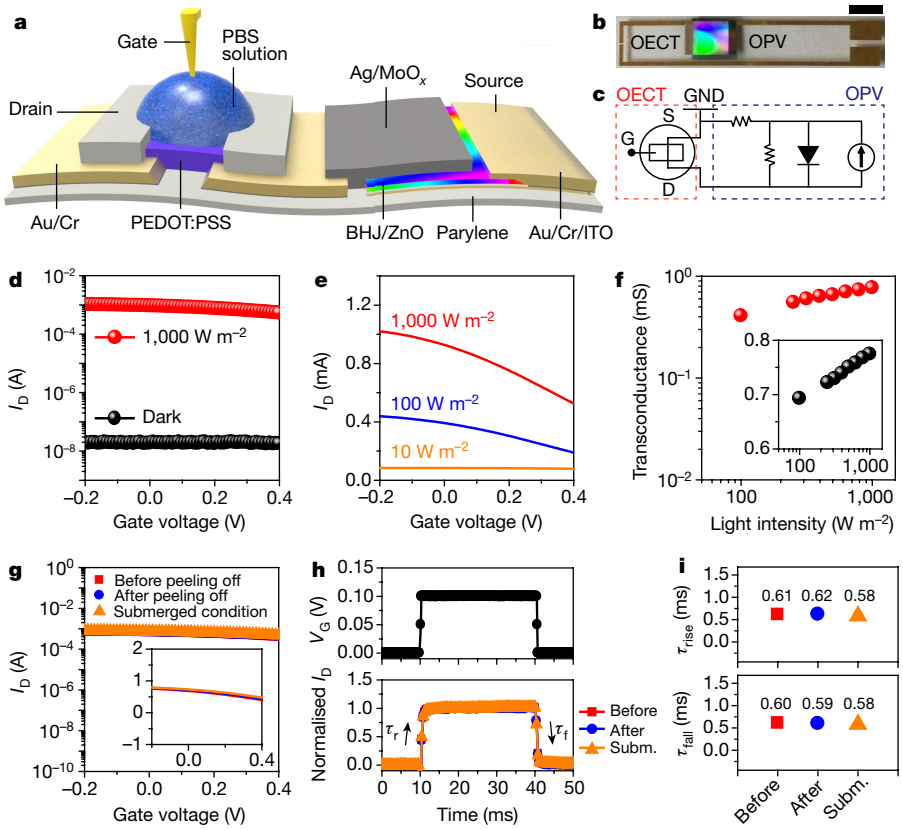


Fig. 3 | Design of the ultra-flexible self-powered integrated device. **a**, Schematic of a double-grating-patterned OPV integrated with an OECT. See Methods for details. **b**, Photograph of the integrated device. Scale bar, 5 mm. **c**, Circuit diagram of the integrated device. D, drain; G, gate; GND, ground; S, source. **d**, I_D–V_G log–linear curve under light illumination (red) and in the dark (black). **e**, I_D–V_G linear–linear curves, showing dependence on the light intensity. **f**, Extracted transconductance values as a function of the light intensity; inset, extracted V_{OC} values (vertical axis, in V) as a function of the light intensity (horizontal axis, in W m^{-2}). **g**, I_D–V_G log–linear curves, showing the effect of the environmental condition (red, before delamination from supporting glass; blue, free-standing condition after delamination; orange, free-standing condition submerged in the physiological PBS solution). Inset shows the data on a linear–linear scale (axes as main figure). **h**, Dynamic I_D output characteristics (bottom) in response to a square wave gate bias, V_G (top). τ_r and τ_f indicate respectively rising time (τ_{rise}) and falling (τ_{fall}) time. **i**, Extracted rising (top, τ_{rise}) and falling (bottom, τ_{fall}) time constants (numbers on symbols), determined under the environmental conditions shown.

performance is achieved because there is no external power line noise, and no ground-loops with the short interconnection wiring[25].

As proof of a self-powered biological sensing application, we demonstrate a conformal self-powered cardiac sensor. Figure 4b displays the wiring diagram. The potential difference between the gel electrode on the chest and the OECT channel on the fingertip (Fig. 4c) acts as the gate bias, affecting the PEDOT:PSS channel conductance. Clear biological signal curves were detected under LED light illumination (approximately 2×10^4 lux) as a result of good skin-conformability. The peak intensity and standard deviation of the recorded cardiac signal were 0.47 μA and 23.5 nA, respectively, resulting in a calculated signal-to-noise ratio of 25.9 dB (Fig. 4d). Our integrated device exhibited threefold higher signal-to-noise ratios than the previous report[21] in which there was high noise influx to both the gate and drain electrodes from external power supplies.

Taking advantage of the mechanical flexibility and bio-compatibility, the fabricated device was attached to the exposed surface of a rat heart (Fig. 4e and Supplementary Video 2). The electrocardiographic (ECG) signals of the rat heart due to the heart-action-evoked potential were measured under an LED lighting condition (approximately 2×10^4 lux). Figure 4f clearly shows that strong ECG signals with an amplitude of 2.96 μA and a standard deviation of 25.2 nA were measured, resulting in a signal-to-noise ratio of 40.02 dB: this value is superior to that in a previous report[20], which recorded a similar frequency of oscillation between epileptiform activity (22.3 dB) using biased OECT devices.

The signal from our device changes when the illumination conditions change. Hence, the LED illumination intensity was maintained constant when we tested our self-powered conformal cardiac sensor. Stabilization of the output power of photovoltaic cells under conditions of fluctuating light intensity is a well-examined topic: such fluctuations of output power can be suppressed using voltage regulators[26] and/or high-pass filters[27]. For example, as the frequency of the ECG can vary[28] from 0.67 Hz to 3 Hz, high-pass filter circuits can reduce slow changes (of less than 0.5 Hz) in the illumination condition. In addition, increasing the effective area of the OPVs and/or the width/length ratio of the OECT also improves the signal-to-noise ratio, so changes in the device dimensions could enable better monitoring under room-light conditions.

Our sensor system comprises two components: an analogue component composed of flexible organic electronics for the detection of biometric signals, and a digital one made of silicon for data transmission. In this work, we have focused on the analogue component in particular, which exhibits extraordinarily stable signals and is powered by ultra-flexible organic photovoltaic cells. Thus, the combination of self-powered analogue and digital components[5,6] offers a way to realize stand-alone, fully self-powered, wearable sensor systems.

Our work can be viewed in the context of previously reported milestones in the development of ultra-flexible OPVs: here, 1D double-grating-patterned high power-per-weight OPVs have (for the first time, to our knowledge) been combined with functional electronic devices to produce self-powered conformal sensors, and the difference in PCE between rigid and flexible OPVs has been overcome. Our approach opens the way to the integration of ultra-flexible organic power sources with functional electronic devices for the precise, sensitive

© 2018 Springer Nature Limited. All rights reserved.

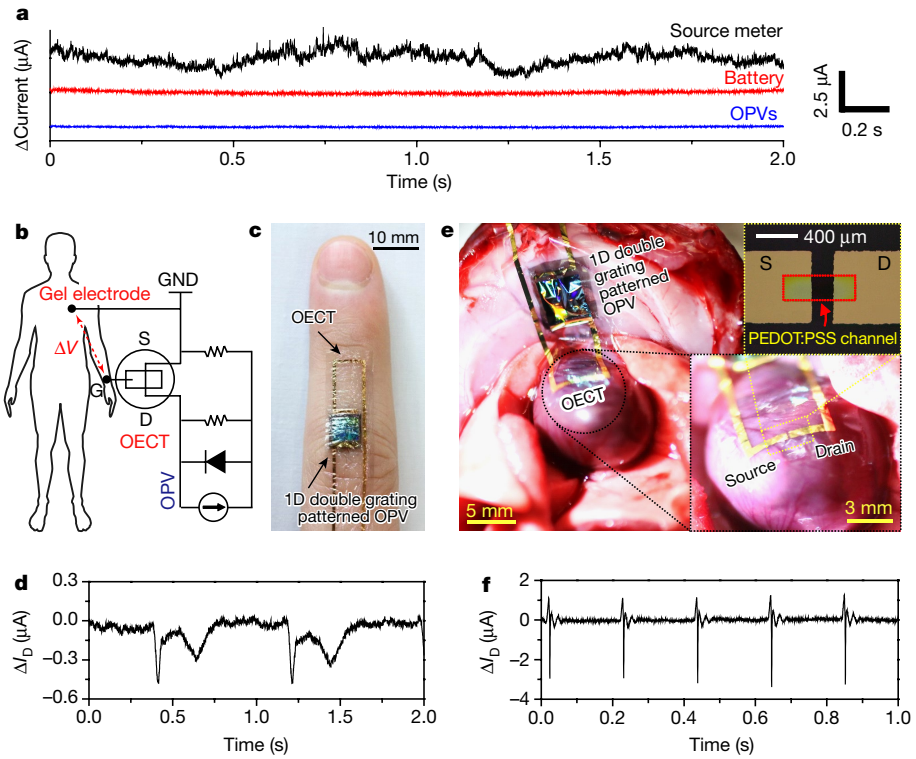


Fig. 4 | Cardiac signal recording by the self-powered ultra-flexible sensor. a, Time traces of the drain current of the OECT, when biased by an external source meter (black), a commercial 1.5 V AA battery (red), and an OPV (blue). The currents are normalized to average background value for comparison. **b**, Wiring diagram for cardiac signal recording. The chest and source electrodes of the integrated device were grounded and the potential difference (ΔV) between the chest and active channel of the OECT functioned as the gate bias. **c**, Photograph of the self-powered integrated electronic device attached to a finger. **d**, Measured output current from the recorded cardiac signal trace, under light illumination. **e**, Photograph of the self-powered integrated electronic device attached to the heart of a rat (left), and enlarged images of the channel area (right, top) and of the source–drain electrode (right, bottom), **f**, Measured output current from the ECG trace, under light illumination.

and continuous data acquisition of biological signals without external power connections.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0536-x.

Received: 24 April 2018; Accepted: 3 August 2018;
Published online 26 September 2018.

1. Li, Z., Zhu, G., Yang, R., Wang, A. C. & Wang, Z. L. Muscle-driven *in vivo* nanogenerator. *Adv. Mater.* **22**, 2534–2537 (2010).
2. Dagdeviren, C. et al. Conformal piezoelectric energy harvesting and storage from motions of the heart, lung, and diaphragm. *Proc. Natl Acad. Sci. USA* **111**, 1927–1932 (2014).
3. Ho, J. S. et al. Wireless power transfer to deep-tissue microimplants. *Proc. Natl Acad. Sci. USA* **111**, 7974–7979 (2014).
4. Song, K. et al. Subdermal flexible solar cell arrays for powering medical electronic implants. *Adv. Healthc. Mater.* **5**, 1572–1580 (2016).
5. Gao, W. et al. Fully integrated wearable sensor arrays for multiplexed *in situ* perspiration analysis. *Nature* **529**, 509–514 (2016).
6. Wang, S. et al. Skin electronics from scalable fabrication of an intrinsically stretchable transistor array. *Nature* **555**, 83–88 (2018).
7. Kim, D.-H. et al. Epidermal electronics. *Science* **333**, 838–843 (2011).
8. Son, D. et al. Multifunctional wearable devices for diagnosis and therapy of movement disorders. *Nat. Nanotechnol.* **9**, 397–404 (2014).
9. Lee, J. W. et al. Soft, thin skin-mounted power management systems and their use in wireless thermography. *Proc. Natl Acad. Sci. USA* **113**, 6131–6136 (2016).
10. Kaltenbrunner, M. et al. Ultrathin and lightweight organic solar cells with high flexibility. *Nat. Commun.* **3**, 770 (2012).
11. Kaltenbrunner, M. et al. Flexible high power-per-weight perovskite solar cells with chromium oxide-metal contacts for improved stability in air. *Nat. Mater.* **14**, 1032–1039 (2015).
12. Jinno, H. et al. Stretchable and waterproof elastomer-coated organic photovoltaics for washable electronic textile applications. *Nat. Energy* **2**, 780–785 (2017).
13. Xu, X., Fukuda, K., Park, S. & Someya, T. Thermally stable, highly efficient, ultraflexible organic photovoltaics. *Proc. Natl Acad. Sci. USA* **115**, 4589–4594 (2018).
14. Heo, S. W. et al. Cumulative gain in organic solar cells by using multiple optical nanopatterns. *J. Mater. Chem. A* **5**, 10347–10354 (2017).
15. Bodurov, I., Yovcheva, T. & Sainov, S. Refractive index investigations of nanoparticles dispersed in water. *J. Phys. Conf. Ser.* **558**, 012062 (2014).
16. Rose, A., Zhu, Z., Madigan, C. F., Swager, T. M. & Bulović, V. Sensitivity gains in chemosensing by lasing action in organic polymers. *Nature* **434**, 876–879 (2005).
17. Yamagiwa, S., Ishida, M. & Kawano, T. Flexible parylene-film optical waveguide arrays. *Appl. Phys. Lett.* **107**, 083502 (2015).
18. Munday, J. N. & Atwater, H. A. Large integrated absorption enhancement in plasmonic solar cells by combining metallic gratings and antireflection coatings. *Nano Lett.* **11**, 2195–2201 (2011).
19. Khodagholy, D. et al. High transconductance organic electrochemical transistors. *Nat. Commun.* **4**, 2133 (2013).
20. Khodagholy, D. et al. *In vivo* recordings of brain activity using organic transistors. *Nat. Commun.* **4**, 1575 (2013).
21. Campana, A., Cramer, T., Simon, D. T., Berggren, M. & Biscarini, F. Electrocardiographic recording with conformable organic electrochemical transistor fabricated on resorbable bioscaffold. *Adv. Mater.* **26**, 3874–3878 (2014).
22. Rivnay, J. et al. High-performance transistors for bioelectronics through tuning of channel thickness. *Sci. Adv.* **1**, e1400251 (2015).
23. Lee, W. et al. Transparent, conformable, active multielectrode array using organic electrochemical transistors. *Proc. Natl Acad. Sci. USA* **114**, 10554–10559 (2017).
24. Northrop, R. B. *Analysis and Application of Analog Electronic Circuits to Biomedical Instrumentation* (CRC, Boca Raton, 2012).
25. Ott, H. W. *Noise Reduction Techniques in Electronic Systems* (Wiley, New York, 1988).
26. Fuketa, H. et al. 16.4 Energy-autonomous fever alarm armband integrating fully flexible solar cells, piezoelectric speaker, temperature detector, and 12V organic complementary FET circuits. In *2015 IEEE Int. Solid-State Circuits Conf. (ISSCC) Digest of Technical Papers* 16.4 1–3 https://doi.org/10.1109/ISSCC.2015.7063043 (IEEE, 2015).
27. Fedotov, A. A. Selection of parameters of bandpass filtering of the ECG signal for heart rhythm monitoring systems. *Biomed. Eng.* **50**, 114–118 (2016).
28. Orphanidou, C. *Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations* (Springer, Cham, 2017).

© 2018 Springer Nature Limited. All rights reserved.

Acknowledgements We thank Toray Industries, Inc., for supplying the polymer material (PBDTTT-OFT) used in our OPVs, and H. Yawo for supplying the optogenetic rat. This work was financially supported by Japan Science and Technology Agency ACCEL grant number JPMJMI17F1.

Reviewer information *Nature* thanks F. Cicoira, D.-H. Kim and K. Leo for their contribution to the peer review of this work.

Author contributions S.P. and T.S. conceived and designed the research. S.P. and S.W.H. fabricated 1D double-grating-patterned ultra-flexible OPV devices and performed electrical and optical measurements. S.P., W.L., T.Y. and M.S. fabricated integrated devices, and performed electrical measurements and the demonstration of ECG monitoring. D.I. and D.H. observed cross-sectional SEM and TEM images. Z.J., K.Y. and H.J. helped in data analysis. S.P., K.F., K.T. and T.S. analysed and interpreted the data, and prepared the manuscript with comments from all the co-authors. T.S. supervised the project.

Competing interests : The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0536-x.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.F., K.T. and T.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Material preparation. ZnO NPs were synthesized by dissolving zinc acetate dehydrate ($Zn(CH_3COO)_2$, Wako, 1.5 g) in methanol (CH_3OH, Junsei, 60 ml), and adding KOH solution (Wako, 99.5%, 0.75 g) under vigorous stirring. After 2 h, the supernatant was removed and washed twice with methanol. The ZnO NPs were centrifuged at 4,000 r.p.m. for 15 min and dissolved in chloroform ($CHCl_3$, Wako, 4 ml). To prevent rapid drying and an irregular ZnO surface, 3 vol.% chlorobenzene (C_6H_5Cl, Wako) was added to the solution before use. PBDTTT-OFT:$PC_{71}BM$ solution was prepared by blending PBDTTT-OFT (Toray Industries) and $PC_{71}BM$ (Solenne BV Corporation, 99.5% purity) at a weight ratio of 0.9:1.1 in a mixed solution of chlorobenzene (C_6H_5Cl, Wako) with 2% 1,8-diiodooctane additive (Sigma-Aldrich) solvent to obtain a total concentration of 20 mg ml^{-1}. For the PEDOT:PSS solution, 20 ml of an aqueous dispersion (Clevios PH 1000) was mixed with 5 wt% ethylene glycol (Sigma-Aldrich) and 0.1 wt% dodecyl benzene sulfonic acid additives (Sigma-Aldrich) to enhance the conductivity and film-forming properties, and 1 wt% (3-glycidyloxypropyl)trimethoxysilane (GOPS) (Sigma-Aldrich) was used to crosslink the film for stable operation in aqueous conditions.

Nanopatterned-PDMS-mould fabrication. The cover layer of a blank DVD-R (TDK) was peeled off with a clean knife to reveal the grating pattern. The DVD-R surface was washed with methanol to remove the dye layer on the grating pattern. The PDMS elastomer (Sylgard 184 base) was prepared by mixing with a curing agent (Sylgard 184 silicone elastomer kit, Dow Corning) at a ratio of 10:1. The elastomers were degassed in a vacuum chamber for 10 min and applied onto the DVD-R surface. After pouring, the sample was degassed again for 10 min. The sample was annealed at 70 °C for 4 h, and the PDMS mould was then peeled off from the surface of the DVD-R mould.

1D double-grating-patterned ultra-flexible OPV fabrication. The glass substrates were sequentially cleaned in 10% neutral detergent (Kanto Kagaku, Cica Clean LX-II), deionized water, acetone, and isopropyl alcohol for 5 min by ultrasonic treatment. After cleaning, oxygen plasma treatment was performed for 10 min to remove the residual chemicals on the glass surface. A fluorinated polymer layer (Novec 1700, 3M) was spin-coated onto the cleaned glass substrate for 60 s at 2,000 r.p.m. and 1-μm-thick parylene film (diX-SR) was deposited by chemical vapour deposition as an ultra-thin substrate. The fluorinated polymer layer is used to decrease the surface energy of the supporting glass, enabling easy delamination of parylene films from the supporting glass substrates. After the deposition of the parylene layer, the substrates were thermally annealed in nitrogen for 1 h at 180 °C. Subsequently, a 500-nm-thick epoxy (SU-8 3005, MicroChem) layer was spin-coated (5,000 r.p.m. for 60 s) for planarization. After the spin-coating process, the SU-8 layer was pre-baked (95 °C for 3 min); this was followed by UV exposure and post-baking (95 °C for 3 min). The annealing processes were performed at 180 °C for 30 min in nitrogen atmosphere, before and after the formation of the SU-8 layer, to minimize unintentional delamination of the ultra-thin substrates in the OPV fabrication process.

A 100-nm-thick layer of bottom ITO electrode was sputtered, without substrate heating, and patterned using photolithography and a wet etching process. Cr/Au electrodes (2.5-nm-thick Cr and 100-nm-thick Au) were deposited as contact pads. A ZnO layer was used as the electron-transporting layer, achieving both band alignment and nano-grating morphology simultaneously. A base ZnO NP layer, as a waveguide layer, was deposited by spin-coating ZnO NP solution at 4,000 r.p.m. for 20 s. Further, the ZnO NP solution was spin-coated again at 3,000 r.p.m. for 7 s, on the first flat ZnO NP layer. After spin coating, the patterned PDMS mould was gently and immediately (within 5 s) placed on top of the ZnO NP layer, and the PDMS mould was delaminated after 60 s. The average groove depth and average period of the patterned ZnO NP layer ranged from 10 to 20 nm and 760 nm, respectively. PBDTTT-OFT:$PC_{71}BM$ solution was deposited in a glove box at 600 r.p.m. for 20 s, resulting in a 150-nm-thick photoactive layer. To form a grating pattern on the photoactive layer, the patterned PDMS mould was placed on the layer, without additional pressure, for 60 s. After peeling off the PDMS mould from the photoactive layer, MoO_x (7.5 nm) and Ag (100 nm, as the top electrode) were deposited sequentially by a thermal evaporator at a pressure below 1×10^{-4} Pa. To complete and form a passivation layer, a 1-μm-thick parylene film was deposited by chemical vapour deposition.

Self-powered conformal cardiac sensor fabrication. ITO electrodes (100-nm thick) were sputtered and patterned via conventional photolithography and a lift-off process on a 1-μm-thick parylene substrate. One side of the ITO was connected to the Cr/Au electrode (2.5-nm-thick Cr and 70-nm-thick Au) as a contact pad. To create and define an open surface on the PEDOT:PSS channel and ITO electrode, a 600-nm-thick parylene layer was additionally deposited and patterned by conventional photo-lithography, using the SU-8 layer and an oxygen reactive ion etching (RIE) process. A solution of PEDOT:PSS was spin-coated at 2,000 r.p.m. for 60 s. After annealing at 100 °C for 1 min, additional annealing was done at 150 °C for 30 min. To define the PEDOT:PSS layer in the channel area of the OECT device, orthogonal patterning and RIE oxygen etching were performed[29]. Photo-resist (OSCoR 5001, Orthogonal) was spin-coated at 1,000 r.p.m. for 30 s and pre-baked at 90 °C for 60 s. UV light was exposed for 10 s through a chrome photomask, using a mask aligner (MA-6, Karl-Suss). After UV exposure, the film was baked at 90 °C for 60 s and developed for 80 s (Orthogonal developer 100, Orthogonal). Covering the channel area of the OECT device, the redundant PEDOT:PSS was etched using a reactive ion etcher (RIE-10NR, Samco) with 30 sccm of oxygen flow at 200 W for 120 s. The Orthogonal photo-resist was removed by a lift-off process with a 3-h duration (Orthogonal stripper 903, Orthogonal). The channel surfaces then were isolated and covered with the PDMS mould, through the OPV fabrication process. Double-grating-patterned photovoltaic devices were fabricated on the ITO electrode, and the top Ag electrode was connected to the drain electrode of the OECT device. The OPVs were passivated by a 1-μm-thick parylene layer. Finally, the PDMS mould was removed from the OECT device. The schematic illustration is depicted in Supplementary Fig. 22.

Compression experiments. The free-standing ultra-flexible OPVs were laminated on a pre-stretched acrylic elastomer substrate (VHB Y-4905J, 3M), before measurement. To record accurate electrical signals, external wiring made of gold was used as the electrical contacts of the free-standing devices. The gold wiring was fabricated by the deposition of 100-nm patterned Au layers through a shadow mask onto 12.5-μm-thick polyimide films. Contact between the devices and the gold wiring was established through electrically conductive adhesive tape (ECATT 9703, 3M). The edges of the gold wiring were connected with alligator clips. The degree of stretching was controlled using a home-made screw machine and a ruler, and the measurements were conducted under ambient conditions. Data obtained from the stretching test were recorded using a Keithley 2400 source meter.

Device characterization and analysis. The current–voltage (I–V) characteristics of the OPVs were recorded under AM 1.5 G (100 mW·cm^{-2}, with the intensity calibrated using a silicon reference solar cell), using a Keithley 2400 source meter under ambient conditions. A typical cell device area of 0.04 cm^2 was defined, using a metal shadow mask. The external quantum efficiency of each device was measured with monochromatic light (SM-250F, Bunkoh-Keiki). OECT characterizations, except for cardiac-signal monitoring, were performed using a solution of PBS as the electrolyte and an Ag/AgCl wire (Warner Instruments) as the gate of the OECT electrode. A droplet of PBS solution was placed on the PEDOT:PSS channel and the Ag/AgCl electrode was dipped in this droplet. The PBS electrolyte was in contact with the PEDOT:PSS channel during measurement. The I–V characteristics and cardiac signal of the OECTs were measured with a semiconductor parameter analyser (B1500A, Keysight). The waveform generator/fast measurement unit (WGFMU) module of the B1500 (Keysight) was used for measuring the drain currents; the other WGFMU module measured the gate voltages of the OECT to determine the time response.

Cardiac-signal monitoring. For the on-skin experiment, the channel of the ultra-flexible integrated device was gently attached onto a wet human fingertip with PBS solution. The source electrodes were grounded with a gel-electrode attached to the human chest. No Ag/AgCl was used for the on-skin experiment. The study protocol was thoroughly reviewed and approved by the ethical committee of the University of Tokyo (approval number KE17-5).

For the in vivo experiments, animal experiments were conducted in accordance with the guidelines of the Animal Experiment Committee at the University of Tokyo. The rat (W-Tg(Thy1-COP4-YFP*)4Jfhy[30], from the National BioResource Project of the Rat in Japan, male, 12 weeks old, 268 g) was anaesthetized using 2–2.5% isoflurane mixed with air, and the skin was incised to expose the surface of the heart. Further, the device was adjusted to contact the anterior surface of the rat's heart. The body fluid present on the surface of the heart was used as an electrolyte for the OECT devices. The source electrodes were grounded by a gel-electrode attached to the sole of the rat's foot. No Ag/AgCl was used for the on-tissue experiment. The ECG data trace was recorded at a constant sampling rate of 10 kHz. The WGFMU module of the B1500 (Keysight) was used to measure the drain currents, without applied bias. The measured data were obtained from a single trace without averaging. Noise filtration above 1 kHz was implemented using the following formula:

$$y_n = \left(\frac{1}{f_{-3\mathrm{dB}}} + t_s\right)\frac{y_{n-1}}{f_{-3\mathrm{dB}}} + \left(\frac{1}{f_{-3\mathrm{dB}}} + t_s\right)\frac{(x_n - x_{n-1})}{f_{-3\mathrm{dB}}}$$

© 2018 Springer Nature Limited. All rights reserved.

where x, t_s, and f_{-3dB} denote the raw data, interval time, and cut-off frequency, respectively. Values of the signal-to-noise ratio (SNR) were calculated as per the following equation:

$$\mathrm{SNR(dB)} = 20\log_{10}(\mathrm{Peak/STD})$$

where 'Peak' and 'STD' denote the peak amplitude and standard deviation of the noise current, respectively.

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

29. Zakhidov, A. A. et al. Orthogonal processing: a new strategy for organic electronics. *Chem. Sci.* **2**, 1178–1182 (2011).
30. Tomita, H. et al. Visual properties of transgenic rats harboring the channelrhodopsin-2 gene regulated by the Thy-1.2 promoter. *PLoS ONE* **4**, e7679 (2009).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0537-9

Multicomponent synthesis of tertiary alkylamines by photocatalytic olefin–hydroaminoalkylation

Aaron Trowbridge[1], Dominik Reich[1] & Matthew J. Gaunt[1]*

There is evidence to suggest that increasing the level of saturation (that is, the number of sp^3-hybridized carbon atoms) of small molecules can increase their likelihood of success in the drug discovery pipeline[1]. Owing to their favourable physical properties, alkylamines have become ubiquitous among pharmaceutical agents, small-molecule biological probes and pre-clinical candidates[2]. Despite their importance, the synthesis of amines is still dominated by two methods: *N*-alkylation and carbonyl reductive amination[3]. Therefore, the increasing demand for saturated polar molecules in drug discovery has continued to drive the development of practical catalytic methods for the synthesis of complex alkylamines[4–7]. In particular, processes that transform accessible feedstocks into sp^3-rich architectures provide a strategic advantage in the synthesis of complex alkylamines. Here we report a multicomponent, reductive photocatalytic technology that combines readily available dialkylamines, carbonyls and alkenes to build architecturally complex and functionally diverse tertiary alkylamines in a single step. This olefin-hydroaminoalkylation process involves a visible-light-mediated reduction of in-situ-generated iminium ions to selectively furnish previously inaccessible alkyl-substituted α-amino radicals, which subsequently react with alkenes to form C(sp^3)–C(sp^3) bonds. The operationally straightforward reaction exhibits broad functional-group tolerance, facilitates the synthesis of drug-like amines that are not readily accessible by other methods and is amenable to late-stage functionalization applications, making it of interest in areas such as pharmaceutical and agrochemical research.

The physiological properties of tertiary aliphatic amines and their ability to interfere with natural neurotransmission pathways have rendered them highly effective pharmaceutical agents[8], in areas that range from the treatment of neurodegenerative disorders (such as Alzheimer's disease)[9] to metabolic syndromes (such as obesity)[10] (Fig. 1a). Traditionally, synthetic routes towards these molecules have required multiple steps and tedious purifications, which severely hamper drug discovery efforts. Therefore, the development of straightforward methods that enable the construction of complex tertiary amines from simple starting materials would have far-reaching implications for both the synthetic and the medicinal chemistry communities. Whereas the abundance, diversity and predictable reactivity of sp^2-hybridized feedstocks has led to the emergence of new transition-metal-catalysed methods of amine synthesis—namely the Buchwald–Hartwig amination[11] and olefin-hydroamination[12–14]—strategies for the synthesis of more complex alkylamines are more limited[13–15]. We reasoned that an operationally simple and mechanistically distinct catalytic process involving available dialkylamines, olefins and aliphatic carbonyl feedstocks would expand the capacity of olefin-hydroaminoalkylation-based strategies to the synthesis of tertiary alkylamines.

We proposed that an electrophilic iminium ion, formed through the well established reaction between secondary alkylamines and alkyl-substituted carbonyls, could be susceptible to a catalytic single-electron reduction process (Fig. 1b). The resulting α-amino radical could then engage a third reaction component, such as an alkene, in a subsequent cross-coupling reaction to form a C(sp^3)–C(sp^3) bond. However, there are few methods that report the generation of 'all-alkyl' α-amino radicals from pre-formed iminium ions, and all have issues relating to their practical application. Subsequently, addition of such α-amino radicals to alkenes has been limited to specific intramolecular examples[16,17]. To overcome these problems we proposed that, first, a visible-light-activated photocatalyst could mediate single-electron transfer (SET) to an in-situ-generated alkyl-substituted iminium ion, generating the desired α-amino radical under mild reaction conditions (Fig. 1b). Second, we considered that a polarity-matched hydrogen-atom transfer (HAT) from a suitable reagent could facilitate cross-coupling to the alkene by intercepting the resulting alkyl-substituted radical. An important feature of this proposed catalytic activation pathway is the regiospecific positioning of the newly formed α-amino radical, made possible by the SET to the iminium ion, regardless of the groups surrounding the reactive centre. Selective formation of an 'all-alkyl' α-amino radical would not be possible via related photocatalytic approaches without the use of pre-functionalized starting materials or inherently selective substrates[18–20] (Fig. 1c).

We were mindful of several factors that could impede the development of the photocatalytic olefin-hydroaminoalkylation process. Whereas protonated imines and iminium ions that are conjugated with multiple aromatic substituents (half-wave reduction potential, $E_{1/2}^{red} = -0.8$ to -1.2 V versus saturated calomel electrode (SCE) in acetonitrile)[21] have been found to partake in a limited range of reductive coupling reactions[22,23], the reduction potential of an 'all-alkyl'-iminium ion could be up to $E_{1/2}^{red} = -1.95$ V (versus SCE in acetonitrile)[21] and would require a highly reducing photocatalyst that may be incompatible with existing functional groups. Moreover, alkyliminium ions are known to exist in (often unfavourable) equilibrium with the corresponding enamines, which can also undergo SET reactions to form radical species, presenting competing pathways[24]. Finally, the addition of α-amino radicals to simple alkenes is known to be low-yielding owing to oligomerization of the resulting radical[25]. We hereby report a comprehensive strategy for the modular and efficient construction of complex tertiary alkylamines via photocatalytic olefin-hydroaminoalkylation.

The initial evaluation of the proposed amine synthesis focused on a reaction between butyraldehyde **1a**, dibenzylamine **2a**, butyl acrylate **3a** and Hantzsch ester **4a** catalysed by $Ir(ppy)_3$ (ppy, 2-phenylpyridinato), under irradiation by visible light (Fig. 2a). Optimized reaction conditions were readily established (for a detailed account of the optimization study, see Supplementary Materials), using 1 mol% $Ir(ppy)_3$ and a 40 W blue light-emitting diode (LED) for 2 hours at room temperature. Under these conditions, near-equimolar quantities of aldehyde, amine and alkene, with 1.5 equivalents of Hantzsch ester in a 0.1 M solution of dichloromethane containing molecular sieves and 20 mol% of propionic acid, formed the desired amine **5a**, which was isolated in 84% yield after chromatography (Fig. 2b). The reaction displayed remarkable selectivity; we observed only trace quantities of the reductive-amination side product, which presumably resulted from HAT to the α-amino radical from **4a**[26].

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. *e-mail: mjg32@cam.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

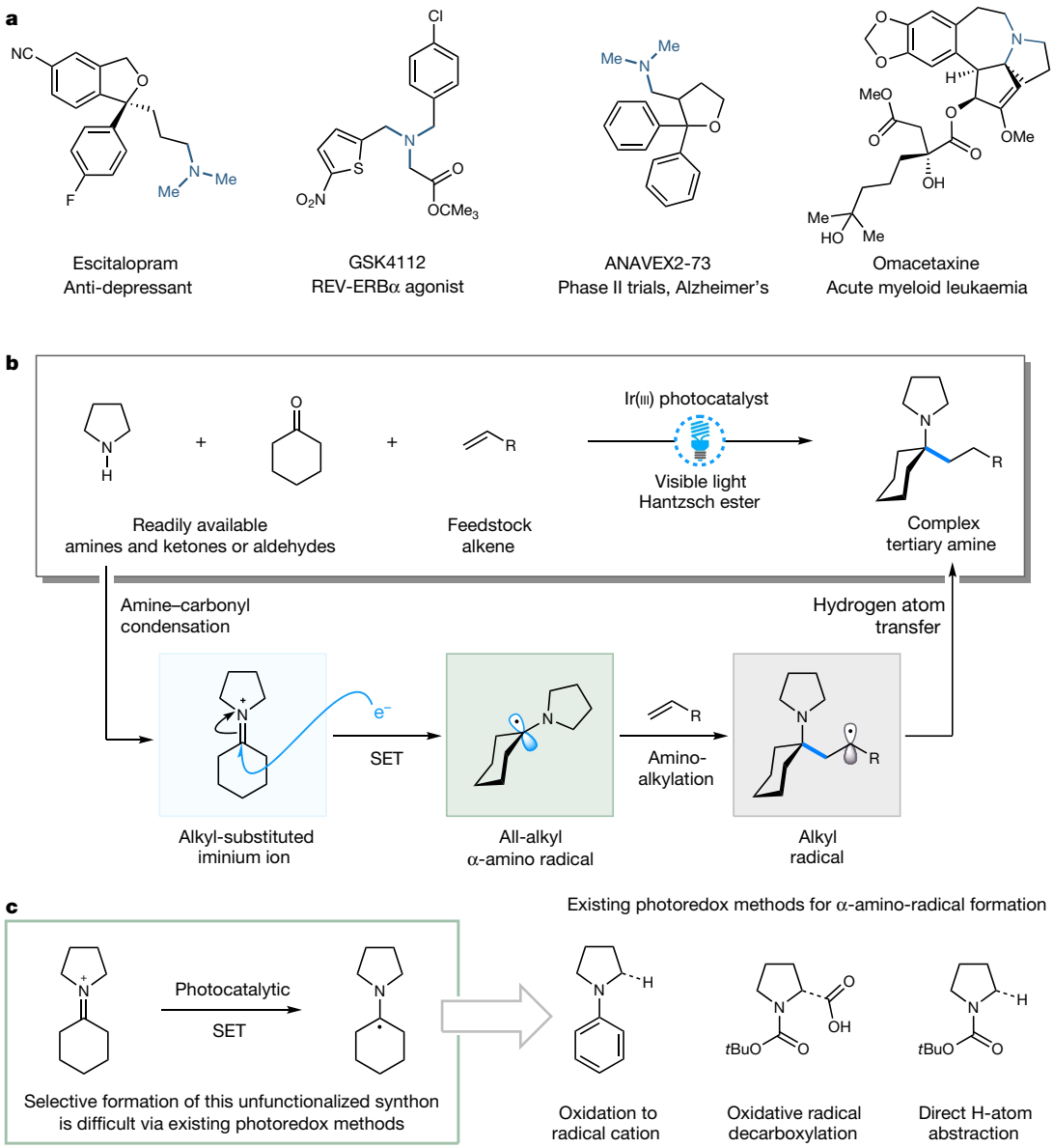


Fig. 1 | Evolution of a strategy for the multicomponent photocatalytic synthesis of tertiary alkylamines. a, Examples highlighting the importance of tertiary amines in the pharmaceutical industry. **b**, Multicomponent photocatalytic olefin-hydroaminoalkylation to yield tertiary alkylamines via the selective formation of alkyl-substituted α-amino radicals. **c**, Catalytic photoredox methods for the generation of α-amino radicals.

Under the optimized conditions we found that various linear aldehydes, including those bearing electron-rich heteroarenes, reacted efficiently to give amines **5a**–**5f** in good yields (Fig. 2b). Aldehydes displaying α-branching produced the expected amines **5g**–**5l** in good yield; notable examples included saturated heterocycles and strained ring features often found in pharmaceutical agents. The α-amino radical formed from formaldehyde and dibenzylamine was effective in the reaction, producing the γ-aminobutyric-acid derivative **5m**; however, benzaldehyde-derived iminium ions failed to deliver the desired cross-coupling products. Next, we found that benzylamine derivatives containing many different functionalized aryl and heteroaryl groups were amenable to the reaction and gave good yields of the corresponding amines **5n**–**5u**. Amines displaying linear and branched alkyl substituents, as well as ester, hydroxyl and nitrile functionality, reacted well to give amines **5v**–**5aa**. The photocatalytic process was effective with a range of electron-deficient alkenes. A reaction with benzyl acrylate **3c** could be adapted to a gram-scale process, yielding 1.4 g of product **5ac** in 84% yield. Notably, acrylonitrile (giving **5ae**) proved a suitable coupling partner, despite being prone to oligomerization in radical reactions. Substituents at either the α- or the β-positions on the alkene (**5ag**–**5ak**) could be accommodated despite the lower electrophilicity and greater steric demand of these acceptors, with a diastereoselectivity of 2.3:1 observed in the reaction of methacrylate to form **5al**. Vinyl pyridine derivatives also proved to be viable acceptors (giving **5am**–**5ap**), enabling facile access to chlorphenamine derivative **5am**. The addition of the α-amino radical to a chiral dehydroalanine derivative led to enantioenriched non-proteogenic amino acid derivative **5aq** in good yield. We also found that perfluorinated alkenes, dienes and electron-deficient alkynes functioned well as acceptors (giving **5ar**–**5at**); the reaction with methyl propiolate delivered (*E*)-allylic amine **5at** in 66% yield as a single geometric isomer.

Our mechanistic proposal for the photocatalytic olefin-hydroaminoalkylation is shown in Fig. 3a. The reaction begins with visible-light excitation of $Ir(ppy)_3$ to the long-lived photoexcited *Ir(III) species, with a lifetime[27], τ, of 1.9 μs. Although this species may be sufficiently reducing (Ir(IV)/*Ir(III), $E_{1/2}^{red} = -1.73$ V versus SCE in acetonitrile)[27] to undergo SET to alkyliminium ion **Int-I**, we recognized that *Ir(III)ppy_3 is efficiently quenched by Hantzsch ester **4a** (*Ir(III)/Ir(II),

© 2018 Springer Nature Limited. All rights reserved.

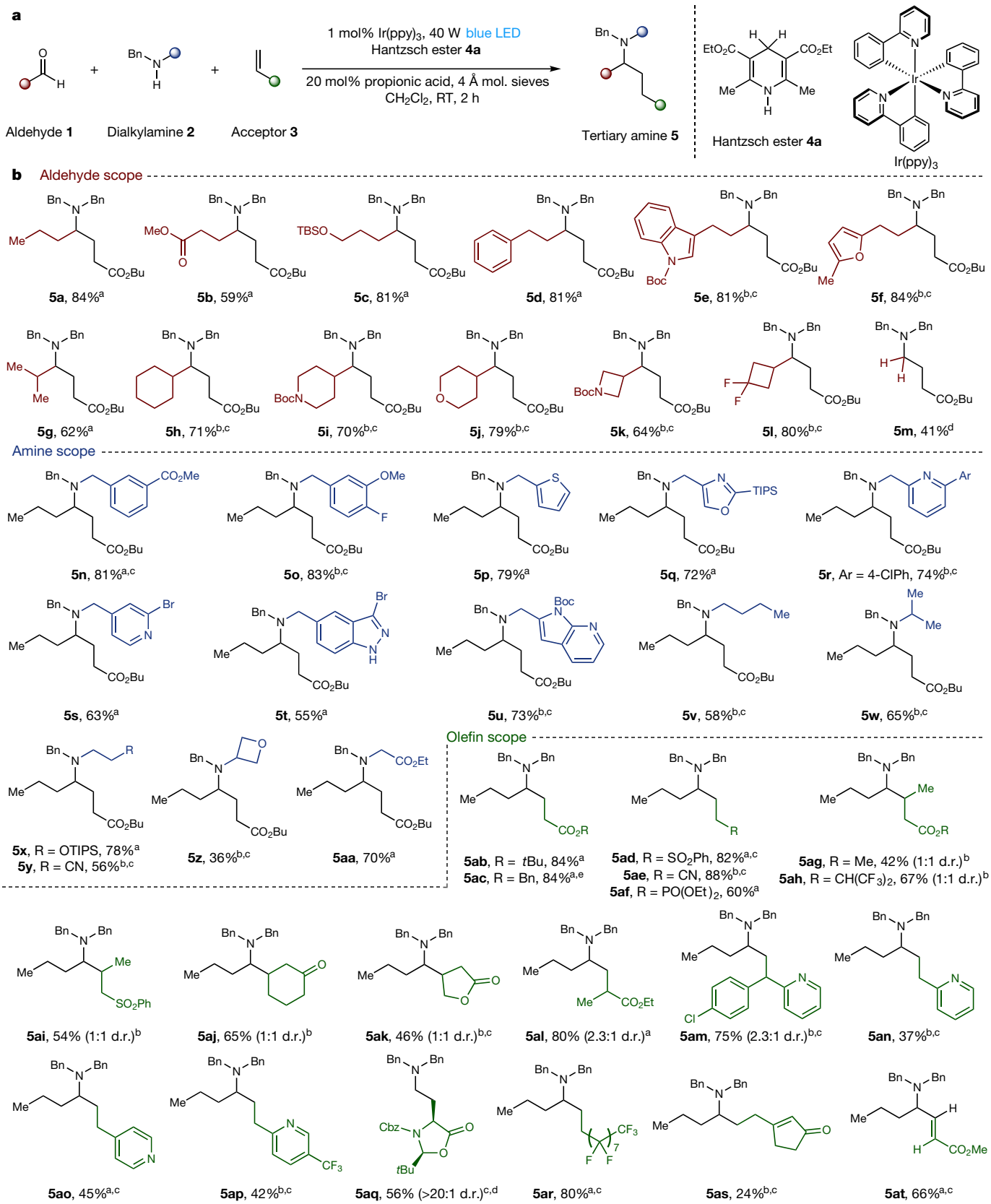


Fig. 2 | Scope of the multicomponent photocatalytic synthesis of tertiary alkylamines. a, Optimized conditions for photocatalytic olefin-hydroaminoalkyation. **b**, Scope of photocatalytic olefin-hydroaminoalkyation. [a]amine:aldehyde:acceptor (1:1.1:1.1), **4a** (1.5 equiv.); [b]amine:aldehyde:acceptor (1:2:2), **4a** (1.5 equiv.); [c]methoxyethyl-Hantzsch ester (1.5 equiv.); [d]paraformaldehyde (5 equiv.), preheated for 1 h; [e]4 mmol scale. Boc, *tert*-butyloxycarbonyl; Cbz, benzyloxycarbonyl; d.r. diastereomeric ratio; ppy, 2-phenylpyridinato; RT, room temperature; TBS, *tert*-butyldimethylsilyl; TIPS, triisopropylsilyl.

© 2018 Springer Nature Limited. All rights reserved.

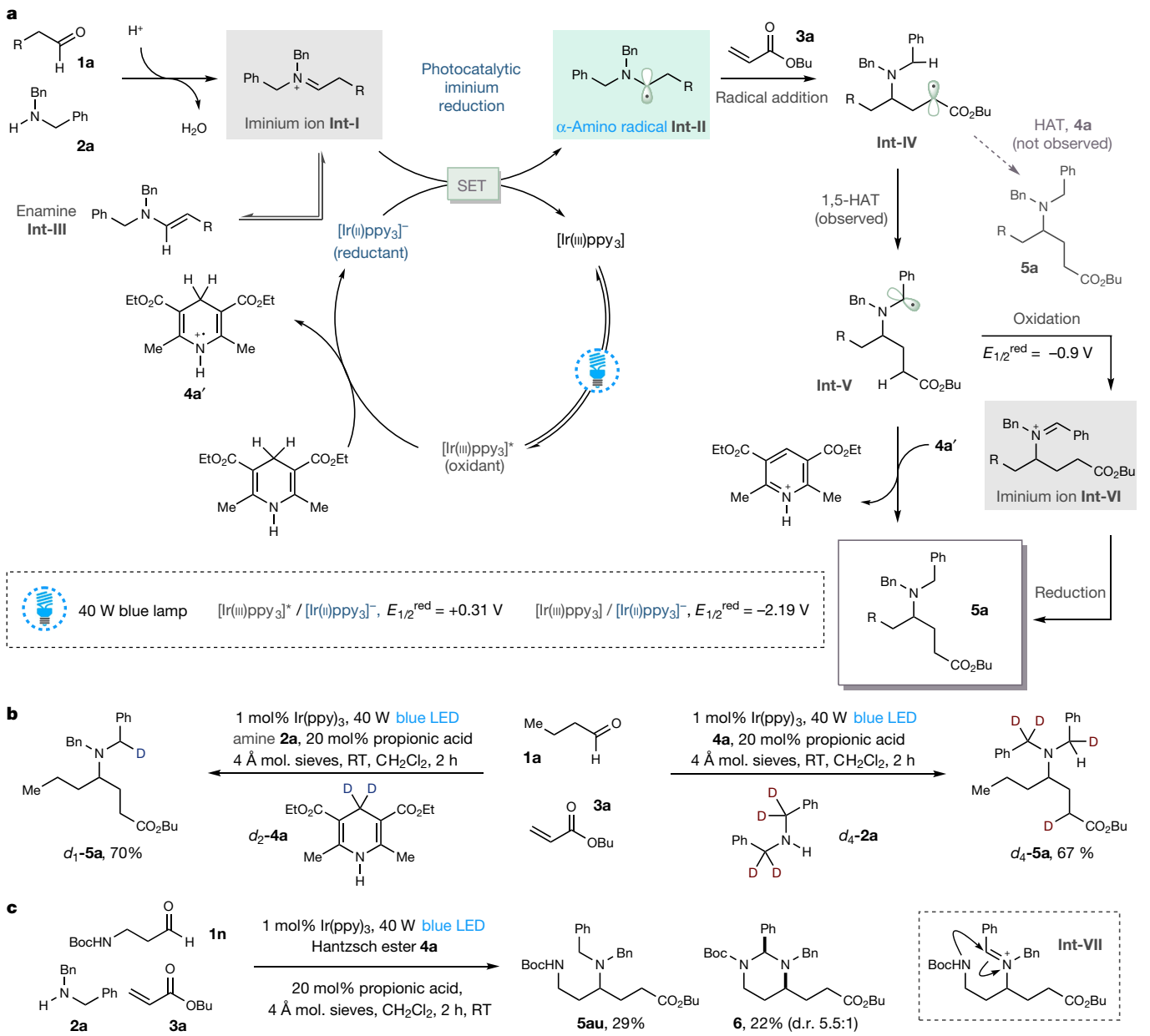


Fig. 3 | Studies towards understanding the mechanism of the multicomponent photocatalytic synthesis of tertiary alkylamines. a, Proposed mechanism for the photocatalytic olefin-hydroaminoalkylation. R = Et. **b**, Deuterium-labelling studies. **c**, Evidence for the iminium-ion redox-relay mechanism.

$E_{1/2}^{red}$ = +0.31 V versus SCE in acetonitrile)[27] leading to $[Ir(II)ppy_3]^-$ and the corresponding Hantzsch ester-radical cation **4a′**. Importantly, $[Ir(II)ppy_3]^-$ is sufficiently reducing (Ir(III)/Ir(II), $E_{1/2}^{red}$ = −2.19 V versus SCE in acetonitrile)[27] to undergo SET with the full range of alkyliminium ions[21], leading to α-amino radical **Int**-II. We identified enamine **Int-III** as the predominant species in the ^{1}H NMR spectrum of the reaction mixture; we believe this is an off-cycle precursor to the iminium ion **Int-I**, with the acid maintaining a low concentration of iminium **Int-I** by protonation of enamine **Int-III**. The α-amino radical **Int-II** now engages the polarity-matched acrylate **3a**, creating a carbon–carbon bond and α-ester radical **Int-IV**. Given the propensity for monosubstituted α-ester radical **Int-IV** to undergo oligomerization[25], we anticipated that an intramolecular 1,5-HAT to the benzylic position may act as a kinetic trap to form stabilized radical **Int-V**[28]. Finally, we expected that the Hantzsch ester **4a** or its radical cation **4a′** would participate in a HAT reaction with **Int-V** to form amine **5a**. A reaction using deuterium-labelled Hantzsch ester d_2-**4a** (Fig. 3b) confirmed our hypothesis; deuterium was incorporated exclusively at the benzylic position of amine d_1-**5a**, showing that 1,5-HAT occurred before interception with **4a** or **4a′**. This theory was further corroborated using labelled dibenzylamine d_4-**2a**, wherein a deuterium atom was transferred to the position adjacent to the ester in amine d_4-**5a**. We also found that an aldehyde bearing a β-nucleophilic group (**1n**) underwent cyclization onto the benzylic position to form **6**, as a side reaction in the formation of **5au** (Fig. 3c). This result suggests that the α-aminobenzyl radical (**Int**-V) can undergo oxidation ($E_{1/2}^{red}$ = −0.9 V versus SCE in acetonitrile)[29] to iminium **Int-VI**, which is accessible to a range of oxidants including *Ir(III)[27]. The selective reduction of benzaldiminium ion **Int-VI** over the initially formed iminium **Int-I** can be rationalized by the inability of **Int-VI** to form a stable enamine intermediate, compared to the interconversion between **Int-I** and **Int-III**. The reduction of **Int-VI** can proceed either by a two-electron process with **4a** or by photocatalytic SET and HAT. Notably, a pathway whereby iminium **Int-I** is translated into a new iminium species **Int-VI** represents an overall mechanism that can be described as a redox relay of iminium ions, which, to the best of our knowledge, has not been reported previously.

© 2018 Springer Nature Limited. All rights reserved.

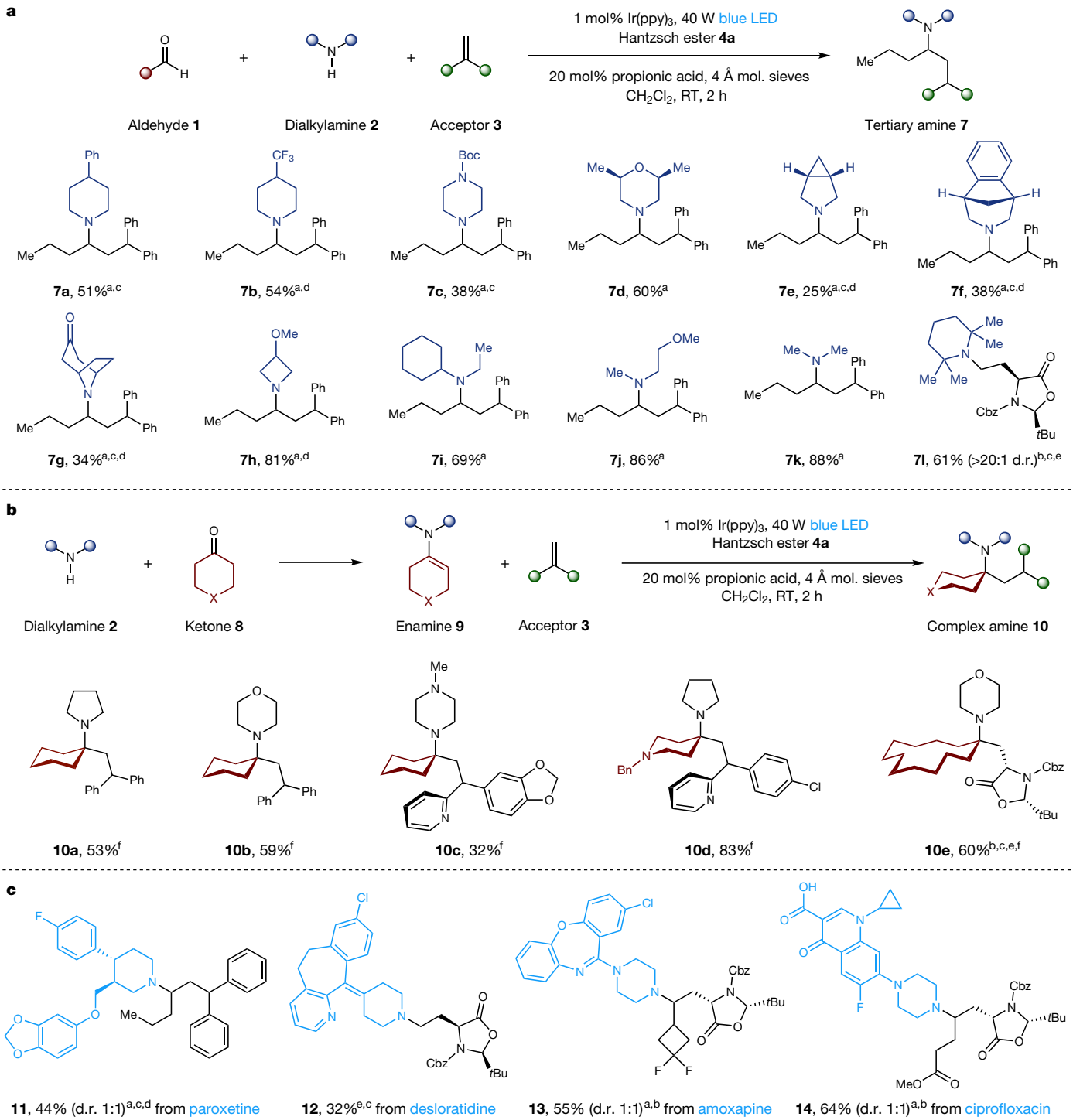


Fig. 4 | Expanding the scope of the multicomponent photocatalytic synthesis of tertiary alkylamines. a, Scope of non-benzylic amines in the photocatalytic olefin-hydroaminoalkylation. **b**, The synthesis of alkyl-substituted α-tertiary amines from dialkylketones. **c**, Late-stage photocatalytic olefin-hydroaminoalkylation with pharmaceutical agents. [a]amine:aldehyde:acceptor (1:2:2), **4a** (1.5 equiv.); [b]acceptor (1.5 equiv.); [c]methoxyethyl-Hantzsch ester (1.5 equiv.); [d]amine hydrochloride and Et_3N (1 equiv.); [e]paraformaldehyde (5 equiv.), preheated for 1 h; [f]enamine (1 equiv.), acceptor (2 equiv.).

In this context, benzylamines not only overcome the inherent challenges posed by the addition of α-amino radicals to electron-deficient alkenes, but also act as a protecting group for primary and secondary alkylamines. Nonetheless, we reasoned that the use of an alkene that would produce a less-electrophilic radical should favour direct reaction with the Hantzsch ester-radical cation **4a′**, obviating the need for the 1,5-HAT process, and hence permitting the use of various dialkylamines. To test this, 1,1-diphenylethylene was combined with aldehyde **1a** and 4-phenylpiperidine under the standard conditions, upon which the desired tertiary amine **7a** was formed in 51% yield (Fig. 4a). Other amine heterocycles were also compatible with this alkene acceptor, giving the amine products **7b**–**7k** in good yield; notably, 1,1-diarylpropylamines are key motifs commonly found in H_1-antihistamines[30]. Among these examples, a number of amines found in pharmaceuticals formed the corresponding complex tertiary amines (**7c**–**7h**), demonstrating the potential to construct 'drug-like' molecules in a single step from readily available materials. The use of non-benzylic amines was not restricted to reactions with 1,1-diarylethenes; the reaction of

© 2018 Springer Nature Limited. All rights reserved.

tetramethylpiperidine, formaldehyde and a dehydroalanine acceptor proceeded smoothly to form amine **7l** as a single diastereomer.

We anticipated that the use of dialkylketones would be an important extension to the photocatalytic process, because these reactions would give rise to α-tertiary amine products[31]. Aware that the formation of ketiminium ions often requires more forcing conditions, we reasoned that protonation of a pre-formed enamine **9** would provide a more accessible source of alkylketiminium ions, as alkylenamines can be readily prepared on a gram scale in one step (Fig. 4b). Under the standard conditions, a range of alkylenamines underwent smooth cross-coupling with alkene acceptors to form α-tertiary amines **10a**–**10e**. The olefin-hydroaminoalkylation method was able to generate complex α-tertiary alkylamine scaffolds in a single step; such scaffolds would be difficult to assemble using classical methods.

Given that dialkylamine motifs are present in a range of small-molecule drugs and pre-clinical candidates, the late-stage functionalization of such molecules would be a powerful demonstration of the utility of this process. To confirm this strategy, we selected four pharmaceutical agents and subjected them to the photocatalytic reaction (Fig. 4c). Each of these architecturally complex amines underwent smooth olefin-hydroaminoalkylation with a variety of aldehydes to furnish the tertiary amine products **11**–**14**. In particular, the combination of desloratidine with formaldehyde and a chiral dehydroalanine acceptor forms a single diastereomer of the tertiary amine derivative **12**, constituting a potentially useful linker strategy through which further functionalization of drug scaffolds could be realized.

We expect that the operational simplicity, efficacy and broad scope of this highly selective, multicomponent photocatalytic amine synthesis will find widespread use among organic chemistry end-users both in academia and in industry. Moreover, we believe that the convenience with which this method generates underexplored alkyl-substituted α-amino radicals will inspire further advances in the synthesis of complex tertiary amine scaffolds.

Data availability

The data that support the findings of this study are available within the paper and its supplementary information files. Raw data are available from the corresponding author on reasonable request. Materials and methods, experimental procedures, useful information, mechanistic studies, optimization studies, ^{1}H NMR spectra, ^{13}C NMR spectra and mass spectrometry data are available in the Supplementary Materials. Crystallographic data are available free of charge from the Cambridge Crystallographic Data Centre (https://www.ccdc.cam.ac.uk/) under reference number CCDC 1819790.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0537-9.

Received: 9 April 2018; Accepted: 10 August 2018;
Published online 26 September 2018.

1. Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: Increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
2. Blakemore, D. C. et al. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
3. Roughley, S. D. & Jordan, A. M. The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **54**, 3451–3479 (2011).
4. Yang, Y., Shi, S.-L., Niu, D., Liu, P. & Buchwald, S. L. Catalytic asymmetric hydroamination of unactivated internal olefins to aliphatic amines. *Science* **349**, 62–66 (2015).
5. Musacchio, A. J. et al. Catalytic intermolecular hydroaminations of unactivated olefins with secondary alkyl amines. *Science* **355**, 727–730 (2017).
6. Johnston, C. P., Smith, R. T., Allmendinger, S. & MacMillan, D. W. C. Metallophotoredox-catalysed sp^3–sp^3 cross-coupling of carboxylic acids with alkyl halides. *Nature* **536**, 322–325 (2016).
7. Matier, C. D., Schwaben, J., Peters, J. C. & Fu, G. C. Copper-catalyzed alkylation of aliphatic amines induced by visible light. *J. Am. Chem. Soc.* **139**, 17707–17710 (2017).
8. Mayol-Llinàs, J., Nelson, A., Farnaby, W. & Ayscough, A. Assessing molecular scaffolds for CNS drug discovery. *Drug Discov. Today* **22**, 965–969 (2017).
9. Lahmy, V., Long, R., Morin, D., Villard, V. & Maurice, T. Mitochondrial protection by the mixed muscarinic/σ_1 ligand ANAVEX2–73, a tetrahydrofuran derivative, in $A\beta_{25-35}$ peptide-injected mice, a nontransgenic Alzheimer's disease model. *Front. Cell. Neurosci.* **8**, 463 (2015).
10. Marciano, D. P. et al. The therapeutic potential of nuclear receptor modulators for treatment of metabolic disorders: PPARγ, RORs, Rev-erbs. *Cell Metab.* **19**, 193–208 (2014).
11. Ruiz-Castillo, P. & Buchwald, S. L. Application of palladium-catalyzed C–N cross-coupling reactions. *Chem. Rev.* **116**, 12564–12649 (2016).
12. Huang, L., Arndt, M., Gooßen, K., Heydt, H. & Gooßen, L. J. Late transition metal-catalyzed hydroamination and hydroamidation. *Chem. Rev.* **115**, 2596–2697 (2015).
13. Kalck, P. & Urrutigoïty, M. Tandem hydroaminomethylation reaction to synthesize amines from alkenes. *Chem. Rev.* **118**, 3833–3861 (2018).
14. Herzon, S. B. & Hartwig, J. F. Hydroaminoalkylation of unactivated olefins with dialkylamines. *J. Am. Chem. Soc.* **130**, 14940–14941 (2008).
15. Perez, F., Oda, S., Geary, L. M. & Krische, M. J. Ruthenium catalyzed transfer hydrogenation for C–C bond formation: hydrohydroxyalkylation and hydroaminoalkylation via reactant redox pairs. *Top. Curr. Chem.* **374**, 35 (2016).
16. Renaud, P. & Giraud, L. 1-Amino- and 1-amidoalkyl radicals: generation and stereoselective reactions. *Synthesis* **1996**, 913–926 (1996).
17. Mariano, P. S. Electron-transfer mechanisms in photochemical transformations of iminium salts. *Acc. Chem. Res.* **16**, 130–137 (1983).
18. Chu, L., Ohta, C., Zuo, Z. & MacMillan, D. W. C. Carboxylic acids as a traceless activation group for conjugate additions: a three-step synthesis of (±)-pregabalin. *J. Am. Chem. Soc.* **136**, 10886–10889 (2014).
19. Nakajima, K., Miyake, Y. & Nishibayashi, Y. Synthetic utilization of α-amino radicals and related species in visible light photoredox catalysis. *Acc. Chem. Res.* **49**, 1946–1956 (2016).
20. Thullen, S. M. & Rovis, T. A mild hydroaminoalkylation of conjugated dienes using a unified cobalt and photoredox catalytic system. *J. Am. Chem. Soc.* **139**, 15504–15508 (2017).
21. Andrieux, C. P. & Savéant, J. M. Electrodimerization: II. Reduction mechanism of immonium cations. *J. Electroanal. Chem.* **26**, 223–235 (1970).
22. Lee, K. N. & Ngai, M.-Y. Recent developments in transition-metal photoredox-catalysed reactions of carbonyl derivatives. *Chem. Commun.* **53**, 13093–13112 (2017).
23. Heinz, C. et al. Ni-catalyzed carbon–carbon bond-forming reductive amination. *J. Am. Chem. Soc.* **140**, 2292–2300 (2018).
24. Capacci, A. G., Malinowski, J. T., McAlpine, M. J., Kuhne, J. & MacMillan, D. W. C. Direct, enantioselective α-alkylation of aldehydes using simple olefins. *Nat. Chem.* **9**, 1073–1077 (2017).
25. Dai, X. et al. The coupling of tertiary amines and acrylate derivatives via visible-light photoredox catalysis. *J. Org. Chem.* **79**, 7212–7219 (2014).
26. Guo, X. & Wenger, O. S. Reductive amination by photoredox catalysis and polarity-matched hydrogen atom transfer. *Angew. Chem. Int. Ed.* **57**, 2469–2473 (2018).
27. Flamigni, L., Barbieri, A., Sabatini, C., Ventura, B. & Barigelletti, F. Photochemistry and photophysics of coordination compounds: iridium. *Top. Curr. Chem.* **281**, 143–203 (2007).
28. Yu, P. et al. Enantioselective C–H bond functionalization triggered by radical trifluoromethylation of unactivated alkene. *Angew. Chem. Int. Ed.* **53**, 11890–11894 (2014).
29. Wayner, D. D. M., Dannenberg, J. J. & Griller, D. Oxidation potentials of α-aminoalkyl radicals: bond dissociation energies for related radical cations. *Chem. Phys. Lett.* **131**, 189–191 (1986).
30. Simons, F. E. R. Advances in H_1-antihistamines. *N. Engl. J. Med.* **351**, 2203–2217 (2004).
31. Hager, A., Vrielink, N., Hager, D., Lefranc, J. & Trauner, D. Synthetic approaches towards alkaloids bearing α-tertiary amines. *Nat. Prod. Rep.* **33**, 491–522 (2016).

Acknowledgements We thank A. Bond for X-ray crystallographic analysis, the EPSRC UK National Mass Spectrometry Facility at Swansea University for HRMS analysis, and M. Nappi, J. C. K. Chu and T. Hunt (AstraZeneca) for useful discussion. We acknowledge the Herschel Smith Scholarship Scheme and AstraZeneca for studentships (to A.T. and D.R., respectively) and the Royal Society for a Wolfson Merit Award (M.J.G.).

Author contributions M.J.G., A.T. and D.R. designed the experiments. A.T. and D.R. performed and analysed the experiments. M.J.G., A.T. and D.R prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0537-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.J.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0532-1

Transience of the North American High Plains landscape and its impact on surface water

Sean D. Willett[1]*, Scott W. McCoy[2] & Helen W. Beeson[2]

Ecosystem diversity and human activity in dry climates depend not just on the magnitude of rainfall, but also on the landscape's ability to retain water. This is illustrated dramatically in the High Plains of North America, where despite the semi-arid modern and past climate, the hydrologic conditions are diverse. Large rivers sourced in the Rocky Mountains cut through elevated plains that exhibit limited river drainage but widespread surface water in the form of ephemeral (seasonal) playa lakes[1], as well as extensive groundwater hosted in the High Plains aquifer of the Ogallala formations[2]. Here we present a model, with supporting evidence, which shows that the High Plains landscape is currently in a transient state, in which the landscape has bifurcated into an older region with an inefficient river network and a younger, more efficient, river channel network that is progressively cannibalizing the older region. The older landscape represents the remnants of the Ogallala sediments that once covered the entirety of the High Plains, forming depositional fans that buried the pre-existing river network and effectively 'repaved' the High Plains[3–6]. Today we are witnessing the establishment of a new river network that is dissecting the landscape, capturing channels and eroding these sediment fans. Through quantitative analysis of the geometry of the river network, we show how network reorganization has resulted in a distinctive pattern of erosion, whereby the largest rivers have incised the older surface, removed fan heads near the Rocky Mountains and eroded the fan toes, but left portions of the central fan surface and the Ogallala sediments largely intact. These preserved fan surfaces have poor surface water drainage, and thus retain ephemeral water for wetlands and groundwater recharge. Our findings suggest that the surface hydrology and associated ecosystems are transient features on million-year timescales, and reflect the mode of landscape evolution.

As central North America emerged from the Western Interior Seaway more than 70 million years ago, it entered a prolonged period of uplift and erosion that continues today, with one exception. From about 20 million to 5 million years (Myr) ago, a wave of sediment from the Rocky Mountains overwhelmed the rivers of the High Plains, depositing silt, sand and gravel, and forming an apron of megafans[7] that completely buried the pre-existing topography. The result was a relatively smooth surface, dipping gently to the east over a distance of more than 1,000 km from the Rocky Mountain front[8,9]. Such transitions between erosion and sediment deposition can result from changes in surface slope, as occur when a continent is tilted dynamically by the underlying mantle flow[10]. The North American mid-continent has been subject to dynamic vertical motion in response to viscous flow driven by the subducting Farallon plate, which induced the formation and destruction of the Western Interior Seaway[10], but is also driving more recent uplift as the drifting North American plate passes over a high-buoyancy region of the upper mantle[11]. Dynamic topography models of this process suggest that a wave of uplift has been propagating eastwards across the southwest United States over the past 30 Myr[11]. This dynamic wave affected the Rocky Mountains first, generating the uplift and erosion that supplied sediment to the High Plains, but as it shifted east, it uplifted and tilted the High Plains, shifting the region from deposition to the erosional state observed today (Fig. 1a). This dynamic tilt was augmented by the isostatic response to erosion[8,9,12,13].

The sediments deposited in the High Plains comprise the geologic formations of the Ogallala Group, which record this uplift event (Fig. 1a). Sediments were deposited in a braided river environment, with coarse sands and gravels as channel deposits, and finer sand and silt as overbank flood deposits[3–5,14]. Mammalian fossils[15–17] and dated volcanic ashes establish that deposition began about 20 Myr ago and that major deposition ended 4 Myr to 5 Myr ago, although sedimentation continued as fluvial valley deposits of limited spatial extent through late Pliocene epoch[12,16,18,19].

The deposition and subsequent incision of the Ogallala sediments resulted in a natural experiment in landscape evolution. Ogallala gravels completely filled the existing river valleys, either forming large alluvial megafans[3] or gently spilling over divides, with a cap of fine-grained eolian (wind-driven) sediments that blanketed valleys and ridges[14]. In either case, the existing erosional channel network was destroyed and replaced with a sediment apron that covers the entire modern High Plains. Today, we are witnessing the establishment of a new river network as the sediment fans are eroded and replaced with an incising, dendritic channel network. The consequence is a transient landscape, split into two components: the older fan surfaces, which largely are not incised, and show low steepness and poorly integrated river channels; and a younger landscape, with a dendritic channel network, that is actively growing in area at the expense of the former.

This ongoing process is evident through examination of geometric characteristics of the river channel network. Scaling between drainage area and channel slope is a near-universal feature of dendritic river channel networks, and variations in this scaling behaviour can be interpreted in terms of network stability[20]. By integrating the inverse drainage area, we obtain a scaling parameter, χ, that characterizes the integrated steepness of a drainage basin and provides a proxy measure of equilibrium elevation (see Methods). We find that large regions of the High Plains are characterized by drainage basins whose headwaters reach very high values of χ. In many cases, these basins have boundaries that are sharply delineated by a strong contrast in χ (Figs. 1b, 2a)—an indication of geometric instability and transience[20]. The basins that exhibit high χ values at relatively low elevation correspond closely with the remnants of the Ogallala formations. The high χ values are a consequence of the unusual geometry of these basins. Rather than exhibiting the dendritic structure that is typical of mature incising or graded channel networks, they show many long, parallel channels with few substantial tributaries and limited branching, commonly defining a radial pattern. These characteristics are observed in modern alluvial fans which, as distributive networks, lack the slope/area scaling relationship of equilibrium channels[21] (Fig. 1b inset and Extended Data Fig. 1). The low degree of branching results in a low rate of downstream accumulation of drainage area and thus a high χ value in the upper reaches of these basins. By contrast, the river basins that cut through or surround the Ogallala remnants are well integrated—with the dendritic

[1]Department of Earth Sciences, ETH, Zurich, Switzerland. [2]Department of Geological Sciences and Engineering, Global Water Center, University Nevada, Reno, NV, USA. *e-mail: swillett@erdw.ethz.ch

© 2018 Springer Nature Limited. All rights reserved.

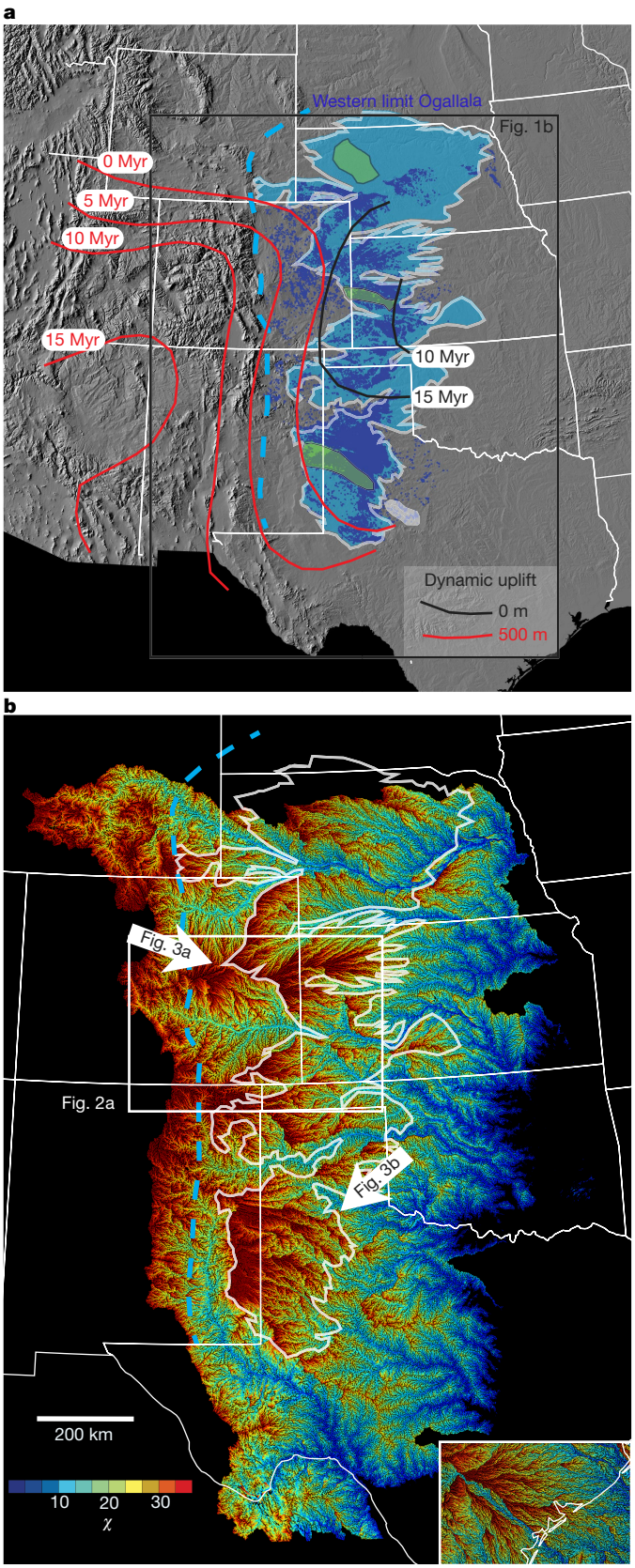


Fig. 1 | River networks, playa lakes and extents of Ogallala sediments across the High Plains. a, Modern (light blue shading) and past (dashed blue line) extents of the Ogallala geologic units. The dark blue points show identified ephemeral playa lakes[32]. The green shading indicates regions of internal drainage. Red lines show the 500-m contours of modelled dynamic topography relative to 30 Myr ago[11]. Black lines represent zero uplift, with subsidence to the east[11]. White lines show US state boundaries. **b**, Map of river channels, coloured according to values of the parameter χ[20] (see Methods). The thick white line shows the extent of the Ogallala Group sediments. For comparison, the inset shows the modern Rio Salado megafan of Argentina at the same scale (see also Extended Data Fig. 1). Channels exhibit radial patterns, emanating from mountain outlets, but also from fan intersection points or avulsion nodes on the surface[33].

pattern that is typical of mature river networks with efficient water and sediment routing—and are characterized by lower values of χ.

Many of the channels mapped on the Ogallala surface are ephemeral or even discontinuous, a point that is partially obscured by our analysis method, which fills small closed depressions in the digital topographic data (see Methods). Particularly in the southern High Plains, there is very little stream flow leaving the Ogallala surface; rather, water accumulates in ephemeral playas, salt lakes or wetlands (Fig. 1a), either returning to the atmosphere through evapotranspiration or entering the groundwater system as recharge to the High Plains Aquifer[22,23]. The impact of high evaporation, infiltration, or a lack of channel integration on our analysis is to increase χ values on the Ogallala surface, decrease erosion rates, and thus increase the contrast between the two landscapes.

The contrast between rivers that reach the Rockies and those that exclusively drain the Ogallala units is also evident in the scaled river profiles (Fig. 2). The profiles of rivers that drain the Ogallala sediments are curved in plots of χ versus elevation, showing neither linearity, nor internal coherence between channels—even those that are closely juxtaposed in space (Fig. 2c and Extended Data Figs. 2–4). These profiles are, however, consistent with rivers flowing directly down a gently dipping, planar surface with limited channel incision and thus low channel concavity[24]. By contrast, rivers that reach the Rocky Mountains have straight to slightly upwardly concave χ-elevation profiles that are more consistent with an equilibrium state (Fig. 2b). The upward concavity appears to be a consequence of incision into the harder rocks of the Rocky Mountains (Extended Data Fig. 5). Many of the large rivers, such as the Arkansas (Fig. 2a), have some tributaries with low steepness and downwardly concave χ profiles. These are tributaries primarily draining the Ogallala units, suggesting that these are regions of the fan surface that have been captured sometime since fan abandonment, but have not yet equilibrated with the trunk channel.

The channel network geometry suggests that the rivers that tap into the Rocky Mountain runoff are incising more quickly and more deeply than the rivers that drain the Ogallala, and are thus cannibalizing the latter through lateral migration of the drainage divide. The topographic asymmetry of water divides would seem to confirm this inference. In fact, the water divides that separate Ogallala-draining rivers from the surrounding basins form some of the most distinct asymmetric escarpments of North America, and include the cliffs of the Llana Estacado in Texas and New Mexico, and the badlands and wooded bluffs from South Dakota to Oklahoma (Fig. 3 and Extended Data Fig. 6)[25]. The escarpments provide independent evidence of landscape transience; their ubiquitous presence along the boundaries between the dendritic and fan-draining channel networks supports our hypothesis that drainage-basin geometry is not in equilibrium, and that the landscape is responding through the replacement of a network of inefficient channels with a new network with higher runoff and erosional efficiency. The escarpments form as the consequence of differential erosion rates and map the transition boundary between the two channel networks.

The Ogallala formations once extended several hundred kilometres east of the modern outcrop boundary[26]. Erosion in this region has occurred by upstream propagation of new river gradients, as channels equilibrate with a base level associated with the Mississippi River. However, the many escarpments on the eastern boundary of the Ogallala outcrop and the predominance of dendritic channel patterns suggest that channel network replacement is also occurring here, competing with, and often outstripping, channel entrenchment[22,26].

© 2018 Springer Nature Limited. All rights reserved.

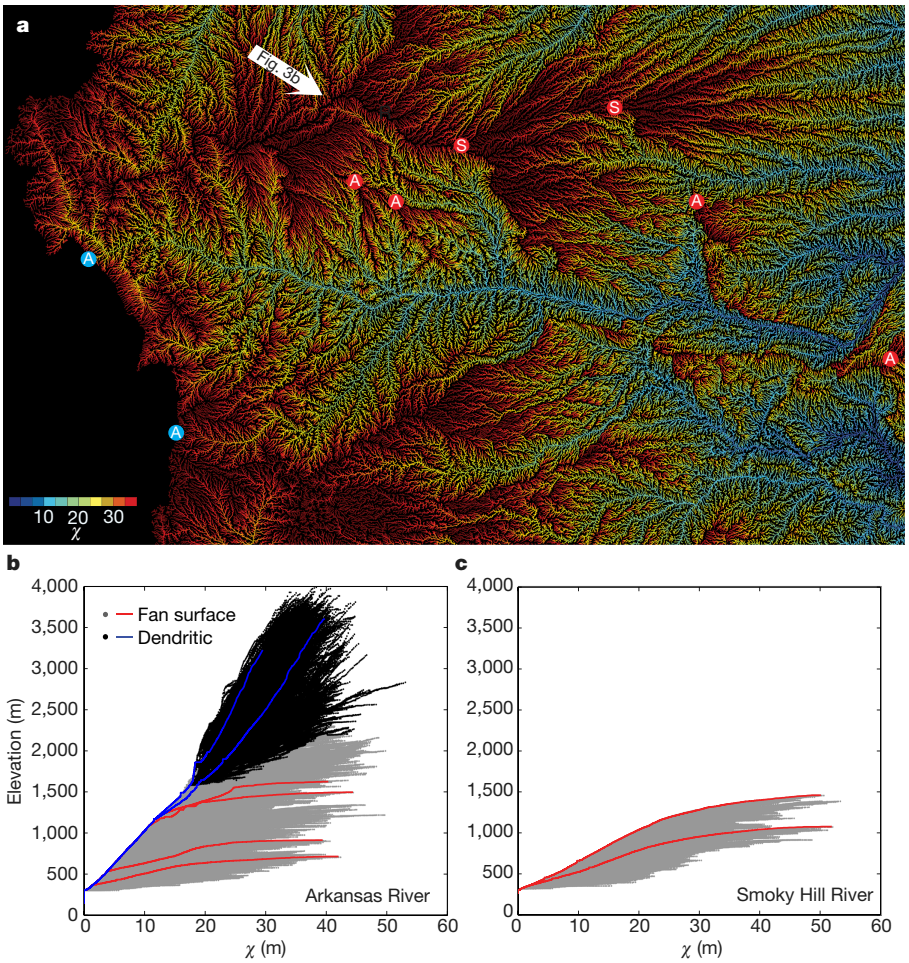


Fig. 2 | Scaled distance/elevation plots for the Arkansas and Smoky Hill river basins. a, χ map centred on the Arkansas River (location shown in Fig. 1b and Extended Data Fig. 2). Circles indicate the heads of channels that are highlighted in **b** and **c**; A, Arkansas; S, Smoky Hill. **b**, χ plots for the Arkansas River. Sets of dendritic and fan-draining rivers define the black and grey envelopes, respectively; blue and red profiles show typical rivers from within each data cloud. **c**, χ plots[27] for the Smoky Hill basin. Major rivers that reach the Rocky Mountains, such as the Arkansas River, show nearly linear χ plots, characteristic of an integrated river. The Smoky Hill River and lateral branches of the Arkansas River show curved, divergent channel profiles, consistent with non-branching rivers that flow directly down a planar surface. Other examples are shown in Extended Data Figs. 3 and 4.

We can compare the rates and time scales of escarpment retreat with the time scale for upstream propagation of channel equilibrium—defined as grading to the base level provided by the Mississippi River—to elucidate why channel network reorganization can predominate as the mode of landscape response. The values of χ shown in Figs. 1 and 2 correspond to the time required to establish a new river grade for an existing channel, provided that we know the rock erodibility parameter, K, and the exponents on the drainage area and slope, m and n (ref. [27]). We estimate these parameters by slope/area analysis of the trunk channels of three of the major rivers, the Platte, Arkansas and Canadian (see Methods and Extended Data Fig. 7); the inferred parameters, combined with the documented incision rate[28], give a minimum equilibration time of 15 Myr to 40 Myr ago for the Republican River and similar Ogallala-draining rivers (Extended Data Fig. 8). We can estimate the rate of escarpment retreat by assuming that it is governed by a disequilibrium river channel that incises according to a slope- and drainage-area-dependent erosion model. The two most common river-erosion models are referred to as detachment limited or transport limited—depending on whether the incision rate is set by the ability to incise bedrock or to move sediment—but both models predict a power-law slope/area scaling and thus are consistent with the χ-mapping analysis presented above. A drainage basin that abuts a moving escarpment (Fig. 3a) grows continuously with time, and thus remains in a transient state. However, if the rock erodibility, climate and total relief remain steady, the slope of an escarpment will adjust so that the water divide maintains a steady velocity. This velocity can be determined analytically for a variety of incision laws (Methods). Using the parameters from the slope/area analysis, we calculated the velocity of the High Plains escarpments for a range of values of the slope exponent, n (Fig. 3c). For an escarpment height of 1 km, small n values, or transport-limited incision, we estimate an escarpment velocity of about 2 mm per year; with a detachment-limited model and larger n values, this goes up quickly to tens of millimetres per year. These velocities imply several tens of kilometres of escarpment retreat since the end of widespread deposition of the Ogallala sediments, consistent with field estimates[26]. To the west, where there is more relief between the fan surface and the incising rivers, escarpment retreat rates should be faster, explaining the removal of the fan heads (Fig. 1b). Overall, we estimate that about 20% of Ogallala deposits, by area, have been eroded by lateral escarpment advance over the past 4 to 5 Myr, outpacing the tens of millions of years required for upstream propagation of the new equilibrium with the Mississippi River.

We propose that this transient model—whereby landscape evolution is dominated by the replacement of an older surface and fluvial network by lateral encroachment of a new channel network—has implications for the evolving hydrology, ecology, biodiversity and even land use of the High Plains. The lack of an integrated channel network on the surface of the Ogallala fan deposits precludes substantial fluvial erosion,

© 2018 Springer Nature Limited. All rights reserved.

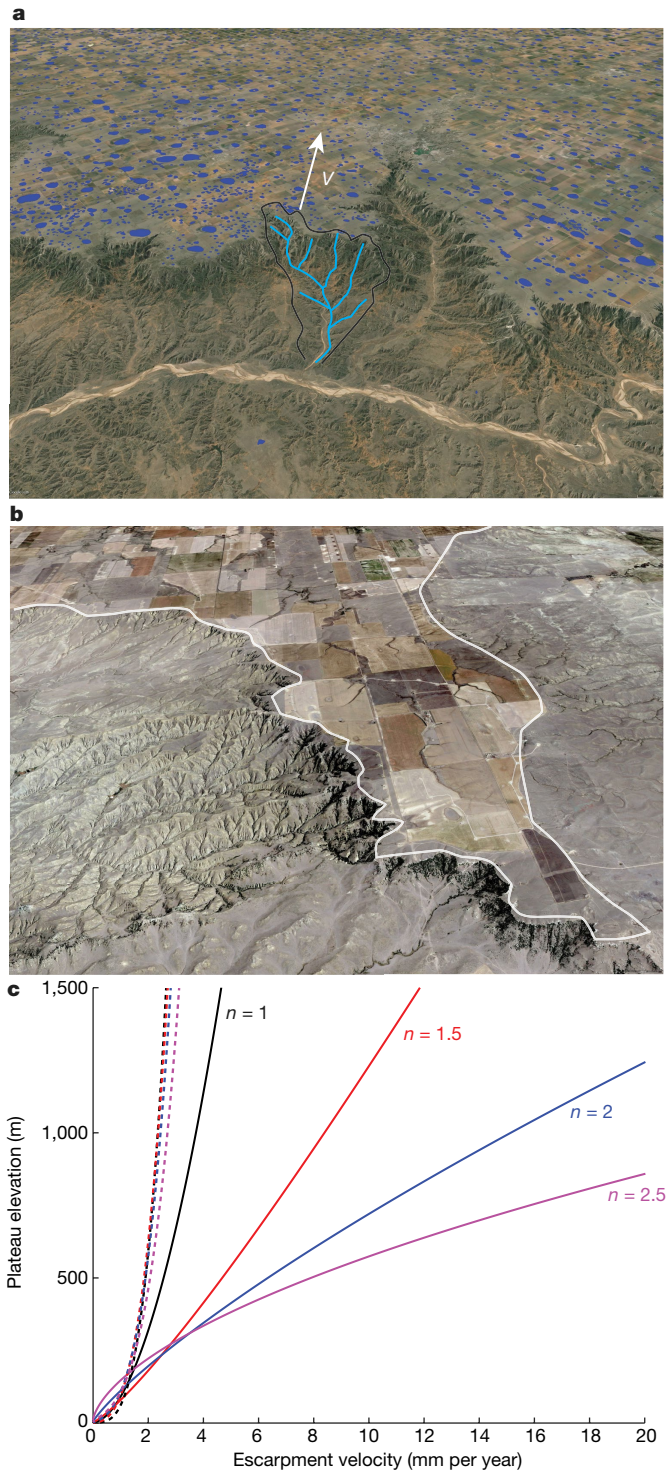


Fig. 3 | Perspective views and predicted velocities of escarpments of the High Plains. a, The escarpment boundary between the Ogallala fan surface and the Red River in the Palo Duro Canyon area of Texas. Playa lakes (blue)[32] on the upper surface and theatre-headed canyons are evidence of groundwater recharge on the surface and discharge and sapping at channel heads[26]. A typical escarpment-draining river is indicated (outlined in black). **b**, Escarpment separating the Ogallala fan surface from the South Platte River (left) and the Arkansas River (right) (location shown in Fig. 2). The white line follows the drainage divide and separates farmland on the Ogallala Group from the dissected badlands. **c**, Escarpment velocity calculated from channel incision models and parameters inferred for the High Plains. Solid lines represent a detachment-limited model for a range of values of the slope exponent, n. Dashed lines represent a transport-limited incision model with the same range of n values (Methods). Constructed from data obtained from Google Earth (https://earth.google.com/web).

and the resulting stability of the landscape has permitted the formation of a thick pedogenic calcrete[16,22], and the preservation of clay-rich loess and thick, well developed soils. Shallow, ephemeral playa lakes form by irregular dissolution of the pedogenic calcrete; these number in the tens of thousands (Fig. 1) and form almost exclusively on the Ogallala deposits (Figs. 1a and 3b)[23]. These playas comprise the most important wetlands of the southern High Plains (Extended Data Fig. 9), providing a bridge for migrating waterfowl, water for pre-modern megafauna, and isolated islands of biodiversity for macroinvertebrates[29,30]. Land use is also affected by the observed mode of landscape evolution, with agricultural activity on the High Plains depending on the mature soils that occur on the Ogallala sediments and the water available from playas or groundwater within the High Plains aquifer system (Extended Data Fig. 10).

Topographic complexity and high biodiversity are often associated with tectonic activity[31], but the landscape evolution model that we propose here for the High Plains suggests that the transition from sediment deposition to fluvial incision can also produce a complex, transient landscape. It seems that the transience in the landscape, and the mode of landscape response by channel network replacement, causes fluvial abandonment of portions of the landscape, where subsequent long-term stability leads to low erosion rates, poor surface drainage, the formation of local wetlands and the production or preservation of loess and soil—conditions that allow a diverse ecological landscape.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0532-1.

Received: 27 October 2016; Accepted: 27 July 2018;
Published online 19 September 2018.

1. Gurdak, J. J. & Roe, C. D. *Recharge Rates and Chemistry Beneath Playas of the High Plains Aquifer—A Literature Review and Synthesis* (US Geol. Survey, Circular 1333, 2009).
2. Gutentag, E. D., Heimes, F. J., Krothe, N. C., Luckey, R. R. & Weeks, J.B., Geohydrology of the High Plains Aquifer in Parts of Colorado, Kansas, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wyoming (US Geol. Survey Prof. Paper 1400-B, 1984).
3. Seni, S. J. *Sand-Body Geometry and Depositional Systems, Ogallala Formation, Texas* (Univ. Texas Austin, Bureau of Economic Geology, Report of Investigations No. 105, 1980).
4. Bretz, J. H. & Horberg, L. The Ogallala formation west of the Llano Estacado. *J. Geol.* **57**, 477–490 (1949).
5. Diffendal, R. F. Regional implications of the geology of the Ogallala Group (upper Tertiary) of southwestern Morrill County, Nebraska, and adjacent areas. *Geol. Soc. Am. Bull.* **93**, 964–976 (1982).
6. Heller, P. L., Dueker, K. & Mcmillan, M. E. Post-Paleozoic alluvial gravel transport as evidence of continental tilting in the U.S. Cordillera. *Geol. Soc. Am. Bull.* **115**, 1122–1132 (2003).
7. Latrubesse, E. M. Large rivers, megafans and other Quaternary avulsive fluvial systems: a potential "who's who" in the geological record. *Earth Sci. Rev.* **146**, 1–30 (2015).
8. Leonard, E. M. Geomorphic and tectonic forcing of late Cenozoic warping of the Colorado piedmont. *Geology* **30**, 595–598 (2002).
9. McMillan, M. E., Angevine, C. L. & Heller, P. L. Postdepositional tilt of the Miocene-Pliocene Ogallala Group on the western Great Plains: evidence of late Cenozoic uplift of the Rocky Mountains. *Geology* **30**, 63–66 (2002).
10. Mitrovica, J. X., Beaumont, C. & Jarvis, G. T. Tilting of continental interiors by the dynamical effects of subduction. *Tectonics* **8**, 1079–1094 (1989).
11. Moucha, R. et al. Deep mantle forces and the uplift of the Colorado Plateau. *Geophys. Res. Lett.* **36**, L19310 (2009).
12. Duller, R. A. et al. Abrupt landscape change post-6 Ma on the central Great Plains, USA. *Geology* **40**, 871–874 (2012).
13. Riihimaki, C. A., Anderson, R. & Safran, E. B. Impact of rock uplift on rates of late Cenozoic Rocky Mountain river incision. *J. Geophys. Res.* **112**, F03S02–F03S15 (2007).
14. Gustavson, T. C. & Winkler, D. A. Depositional facies of the Miocene-Pliocene Ogallala Formation, northwestern Texas and eastern. *N. M. Geol.* **16**, 203–206 (1988).
15. Skinner, M. F. & Johnson, F. W. Tertiary stratigraphy and the Frick Collection of fossil vertebrates from north-central Nebraska. *Bull. Am. Mus. Nat. Hist.* **178**, 215–368 (1984).
16. Ludvigson, G. A., Sawin, R. S., Franseen, E. K. & Watney, W. L. A review of the stratigraphy of the Ogallala Formation and revision of Neogene (Tertiary) nomenclature in Kansas. *Curr. Res. Earth Sci.* bulletin 256, part 2 (Kansas Geological Society, 2009).

© 2018 Springer Nature Limited. All rights reserved.

17. Winkler, D. A. *Stratigraphy, Vertebrate Paleontology and Depositional History of the Ogallala Group in Blanco and Yellowhouse canyons, Northwestern Texas*. PhD thesis, Univ. Texas, Austin (1985).
18. Wisniewski, P. A. & Pazzaglia, F. J. Epeirogenic controls on Canadian River incision and landscape evolution, Great Plains of Northeastern New Mexico. *J. Geol.* **110**, 437–456 (2002).
19. Lazear, G., Karlstrom, K., Aslan, A. & Kelley, S. Denudation and flexural isostatic response of the Colorado Plateau and southern Rocky Mountains region since 10 Ma. *Geosphere* **9**, 792–814 (2013).
20. Willett, S. D., McCoy, S. W., Perron, J. T., Goren, L. & Chen, C. Y. Dynamic reorganization of river basins. *Science* **343**, 1248765 (2014).
21. Horton, B. K. & DeCelles, P. G. Modern and ancient fluvial megafans in the foreland basin system of the central Andes, southern Bolivia: implications for drainage network evolution in fold-thrust belts. *Basin Res.* **13**, 43–63 (2001).
22. Wood, W. W. Role of ground water in geomorphology, geology, and paleoclimate of the Southern High Plains, USA. *Ground Water* **40**, 438–447 (2002).
23. Osterkamp, W. R. & Wood, W. W. Playa-lake basins on the Southern High Plains of Texas and New Mexico: part I. Hydrologic, geomorphic, and geologic evidence for their development. *Geol. Soc. Am. Bull.* **99**, 215–223 (1987).
24. Zaprowski, B. J., Pazzaglia, F. J. & Evenson, E. B. Climatic influences on profile concavity and river incision. *J. Geophys. Res.* **110**, F03004 (2005).
25. Wells, P. V. Scarp woodlands, transported grassland soils, and concept of grassland climate in the Great Plains region. *Science* **148**, 246–249 (1965).
26. Higgins, C. G. & Osterkamp, W. R. in *Groundwater Geomorphology: The Role of Subsurface Water in Earth-Surface Processes and Landforms* (eds Higgins, C. G. & Coates, D. R.) Ch.14 (Geol. Soc. Am. Spec. Pap. 252, 1990).
27. Perron, J. T. & Royden, L. An integral approach to bedrock river profile analysis. *Earth Surf. Process. Landforms* **38**, 570–576 (2012).
28. Dethier, D. P. Pleistocene incision rates in the western United States calibrated using Lava Creek B tephra. *Geology* **29**, 783–786 (2001).
29. Bolen, E. G., Smith, L. M. & Schramm, H. L. Playa lakes: prairie wetlands of the Southern High Plains. *Bioscience* **39**, 615–623 (1989).
30. Hall, D. L., Willig, M. R., Moorhead, D. L. & Sites, R. W. Aquatic macroinvertebrate diversity of playa wetlands: the role of landscape and island biogeographic characteristics. *Wetlands* **24**, 77–91 (2004).
31. Badgley, C. et al. Biodiversity and topographic complexity: modern and geohistorical perspectives. *Trends Ecol. Evol.* **32**, 211–226 (2017).
32. Playa Lakes Joint Venture Project On-line data set. *Probable Playas Version 4* http://pljv.org (2015).
33. Ganti, V., Chu, Z., Lamb, M. P., Nittrouer, J. A. & Parker, G. Testing morphodynamic controls on the location and frequency of river avulsions on fans versus deltas: Huanghe (Yellow River), China. *Geophys. Res. Lett.* **41**, 7882–7890 (2014).

Acknowledgements We thank the Playa Lakes Joint Venture Project for data files on playa locations. We also thank J. K. Caves Rugenstein for discussions and suggestions, particularly with reference to birds. H.W.B. was supported in part by the National Aeronautics and Space Administration under grant NNX15AIO2H.

Reviewer information *Nature* thanks R. Duller and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.D.W. and S.W.M. conceived the project. S.D.W. provided background information and was responsible for hydrologic data and analytical theory. H.W.B. and S.W.M. constructed χ maps, χ plots and land-use maps, and conducted concavity and relief analysis. All authors contributed to the writing of the manuscript.

Competing interests : The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0532-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.D.W.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Playa identification. The locations of playa lakes shown in Figs. 1, 3 are from a compilation by the Playa Lakes Joint Venture, available at www.pljv.org[32]. These data were obtained primarily from aerial or satellite imagery and were compiled from a variety of sources and organizations, including the Natural Resources Conservation Service, Ducks Unlimited, the National Wetlands Inventory, the National Agricultural Imagery Program, and the National Hydrography Data set of the US Geological Survey. More than 75,000 probable playas have been identified and mapped across the states of Texas, Oklahoma, Kansas, Colorado, New Mexico and Nebraska, but large gaps in coverage remain, so maps should not be regarded as comprehensive.

χ maps and profiles. For any incision model that results in power-law scaling between channel slope and drainage area, an equivalent relationship exists relating the elevation of a point in the river network, z, at distance x along a river channel and the integral quantity, χ:

$$z(x) = z_b + K_s A_0^{-\theta}\chi$$

where z_b is the elevation of a reference base level, K_s is the proportionality constant, and χ is defined as:

$$\chi = \int_{x_b}^{x}\left(\frac{A_0}{A(x')}\right)^{\theta} dx'$$

where A is the drainage area at x, base level for the river is at x_b, θ is the negative of the slope of the logarithmic scaling relationship between slope and area, sometimes referred to as concavity, and A_0 is an arbitrary scaling area, used to give χ dimensions of length. With both detachment-limited and transport-limited fluvial incision models, this describes a linear relationship at steady state given constant, homogeneous parameters and uplift rate. The relationship between the proportionality constant, K_s, and fluvial incision parameters is discussed below in the context of the slope/area analysis.

For χ maps and profiles, we calculated flow directions, flow paths, and accumulated flow (upstream drainage area) from a digital elevation model. For flow direction and paths, closed basins were filled, and steepest descent neighbours found for local flow direction. We used a critical drainage area, A_c, equal to 0.5 km^2 as a limit to the upstream extent of the fluvial channel network. This parameter limits the upstream extent of our analysis, but does not affect the downstream value of χ. We used a 90-m-resolution digital elevation model produced from the Shuttle Radar Topography Mission (SRTM) and downloaded from Open Topography (http://www.opentopography.org/). We used a scaling area, A_0, of 1 m^2 and a base level of 294 m, which is the elevation of the confluence between the Platte and Missouri rivers. We set θ to 0.45 for the main analysis, but investigated maps using a range of θ. First-order features in the χ maps do not appear substantially different in the range of 0.3 to 0.45 (compare Fig. 1 with Extended Data Fig. 2). Note that more common protocol determines θ through optimization of linearity or minimization of variance in the resultant χ-elevation plots for the entire landscape in question. However, this methodology is appropriate only for equilibrium or near-equilibrium river profiles. For the highly transient profiles of our study area, these methods would predict θ values of close to zero for drainages on fans capped by Ogallala sediments (Extended Data Fig. 5).

Slope/area analysis and estimation of *K*. Slope/area scaling was evaluated for several of the trunk channels crossing the High Plains in order to estimate K and concavity, θ. From the regression of log(slope) against log(area), we obtain a concavity (θ) and a channel steepness, K_s. It is necessary to assume one exponent in order to translate these into the parameters K, m and n for a detachment-limited model:

$$K = \frac{U}{K_s^n}, m = n\theta$$

and for a transport-limited model:

$$K = \frac{U}{K_s^n}, m = n\theta + 1$$

We made logarithmic regressions of slope and area for three of the largest rivers of the High Plains—the South Platte, the Arkansas and the Canadian—masking regions that drain the Ogallala surfaces. We found a consistent concavity of 0.3 and steepness of about 1.5 (Extended Data Fig. 7). These were used with an estimated incision rate of 0.05–0.1 mm per year (ref. [28]) and a range of n values to obtain parameters for basin response times and the escarpment retreat velocity.

Response time for a river basin. The time taken for a transient base-level change to travel up a river is defined as:

$$\tau = \int_0^x \frac{1}{KA^m S^{n-1}} dx$$

The nonlinear dependence on slope means that response times cannot be calculated for general n, but a scale analysis can be made assuming that the profile is close to equilibrium with a characteristic incision rate, U. The slope can then be approximated as:

$$S \approx \left(\frac{U}{K}\right)^{\frac{1}{n}} A^{-\frac{m}{n}}$$

giving:

$$\tau = \frac{U^{\frac{1}{n}-1}}{K^{\frac{1}{n}}}\chi = \frac{K_s}{U}\chi$$

Using the derived K_s of 1.5, θ of 0.3 and assumed uplift rate of 0.05–0.1 mm per year implies that a river of length χ = 1,200 m has a response time of 18 to 36 Myr.

Relief maps. We calculated relief on the High Plains for each grid cell as the maximum elevation minus minimum elevation within a moving window with a radius of 200 m, using the 90-m-resolution digital elevation model described above.

Wetland designation. We visually compared the distribution of wetland types on and off the Ogallala sediments by using geospatial data on the characteristics of wetlands as classified by the US Fish and Wildlife Service National Wetlands Inventory[34], a publicly available data set on the USA's wetland resources.

Land-use calculations. We mapped land cover using the 2015 National Land Cover Database for the conterminous United States[35]. We calculated the percentages of land classified as 'Cultivated crops' within the extent of the Ogallala sediments, excluding the North Platte River basin where aeolian deposits prevent cultivation. We compared this value with land classified as 'Cultivated crops' outside of the Ogallala sediments boundary but within the north–south extent of the Ogallala sediments, east of the Rocky Mountain front and west of the 600 mm per year precipitation line (Extended Data Fig. 8), where it appears that additional precipitation eliminates the importance of the Ogallala aquifer as a source of water. Other work has also noted the correspondence between Ogallala presence and soil type and thickness[36].

Analytical solutions for water-divide migration rate. A river basin abutting a non-eroding, flat highland or plateau and incising downward will create an escarpment capped by a water divide (Fig. 3). Incision will drive this water divide into the elevated plateau with time. In the absence of additional uplift—tectonic or isostatic—the total relief across the drainage basin will remain constant, and the lower reaches of the basin will largely stay in equilibrium with the base level established at its outlet. If the channel gradient in these equilibrium reaches is low, as would be the case with no uplift, the basin relief will be concentrated in the upper reaches and hill slopes. Given constant relief and constant rock and climatic properties, we postulate that the rate of migration of the water divide will reach a constant value and hold that velocity as the escarpment migrates inland, a result also obtained in ref. [37] for a more restrictive case. The upper reaches draining the escarpment will always be in disequilibrium—that is, steeper than would be predicted for a river in equilibrium with its base-level incision rate. The system will evolve until the river profile, the incision rate (which varies with distance along the channel) and the rate of retreat of the water divide reach a dynamic equilibrium. We consider two incision laws—a detachment-limited and a transport-limited case—and solve for this dynamic profile and the velocity of the divide.

Detachment-limited river. For a river channel that is incising according to a detachment-limited, stream power model, a river profile responding to an uplift rate, U, is described by:

$$\frac{dz(x,t)}{dt} = U - KA(x,t)^m\left(-\frac{dz(x,t)}{dx}\right)^n$$

where z is channel elevation, A is the drainage area, m and n are empirical, rational exponents, and K is an erodibility parameter including effects of rock type and runoff, as discussed in the slope/area analysis. We choose positive x values as 'downhill' so that the slope is always negative, and retain a negative sign to assure that the nth power of slope is always positive. This equation is subject to a fixed boundary condition at its base level, but a moving boundary condition at its water divide, which affects not only the drainage area, A, but also the length of the domain over which this equation is valid. We can make some simplifying assumptions to make the problem tractable. First, we assume that Hack's Law holds; this law relates the length of the longest channel in a basin to the drainage area, as follows: $A = k_a x^h$, with x measured from the water divide, and h being an empirical constant. In the subsequent analysis we drop k_a by including it in K. We assume that Hack's Law holds at all times, which implies that as the basin grows headward into the plateau, it also widens. We assume that uplift

© 2018 Springer Nature Limited. All rights reserved.

is negligible, that is, $U=0$. Under these conditions, the divide velocity will attain a constant value and we can define the reference frame for the problem to be fixed to the water divide, so that $x=0$ at the divide. In this reference frame, the boundaries and domain are fixed, but material (the rock underlying the channel) is moving. This formulation thus adds an advective term to the problem, but eliminates the time derivative, so that the governing equation for the steady-state profile is:

$$-v\frac{\mathrm{d}z}{\mathrm{d}x}-Kx^{hm}\left(-\frac{\mathrm{d}z}{\mathrm{d}x}\right)^{n}=0 \tag{1}$$

This equation is solvable by integration, taking a boundary condition of fixed elevation at an point where the channel makes the transition to hillslope processes, defined to be at $x=x_c$ and $z=z_c$:

$$z=\begin{cases} z_c+\dfrac{1}{(1-hmn-1)}\left(\dfrac{v}{K}\right)^{\frac{1}{n-1}}\left(x_c^{\,1-\frac{hm}{n-1}}-x^{1-\frac{hm}{n-1}}\right) & \text{for } \dfrac{hm}{n-1}\neq 1, n\neq 1 \\ z_c+\left(\dfrac{v}{K}\right)^{\frac{1}{n-1}}\ln\left(\dfrac{x_c}{x}\right) & \text{for } \dfrac{hm}{n-1}=1, n\neq 1 \end{cases} \tag{2}$$

There is no non-trivial solution for the case of $n=1$, although we will present a solution for the limiting case below. It is also interesting that, as a first-order wave equation, there is a characteristic solution to the problem with the celerity of information along the characteristic as C, defined by:

$$C=Kx^{hm}\left(\frac{\mathrm{d}z}{\mathrm{d}x}\right)^{n-1}$$

For the solution in equation (2), C is constant with x and equal to v. The solution implies a scaling relationship between slope and area as:

$$S=-\left(\frac{v}{K}\right)^{\frac{1}{n-1}}A^{-\frac{m}{n-1}}$$

The full profile can be obtained by selecting a hillslope process to constrain x_c and z_c. A diffusion model with an advection term could be used[38], but a simpler result is obtained by assuming that the hillslope takes a constant slope at a threshold value and that the river profile maintains slope continuity at the transition. In the advection problem, slope continuity is equivalent to incision-rate continuity. For a critical threshold angle, θ, the transition to the fluvial channel is defined to be at:

$$\begin{aligned} x_c&=(\tan\theta)^{-\frac{n-1}{hm}}\left(\frac{v}{K}\right)^{\frac{1}{hm}} \\ z_c&=z_p-x_c\tan(\theta) \end{aligned} \tag{3}$$

where z_p is the elevation of the plateau. Equations (2) and (3) define the full profile for a given set of physical parameters including v, but we expect v to be emergent from the solution, in which case we need an additional constraint. This is provided by the length of the channel, from divide to an effective base level, which we refer to as x_0. Given our assumption of steady state, this is implicitly assuming that x_0 is moving at velocity v with the water divide and maintains an elevation of $z=0$. We regard the point x_0 as being at the transition between the equilibrium reach and the disequilibrium reach, as defined in the fixed reference frame. The assumption of $z=0$ at x_0 is strictly true only for $U=0$, but provided that the channel slope of the equilibrium reach is small compared to the disequilibrium reach, the total fluvial relief at x_0 will be small and the assumptions that $U=0$ and $z(x_0)=0$ are justified. We acknowledge that, in nature, the elevation of this point will be rising slowly with time as a zero-slope channel is not physical. It is not clear how to pick x_0, but solutions are not sensitive to this parameter for most values of the exponents in the erosion model. Plugging in the condition $x=x_0$, $z=0$, we obtain an expression that can be solved for v. Velocity cannot be obtained explicitly, so it is convenient to write this in terms of the plateau elevation to give a parametric solution for v:

$$z_p=\begin{cases} x_c\tan\theta-\dfrac{1}{1-\frac{hm}{n-1}}\left(\dfrac{v}{K}\right)^{\frac{1}{n-1}}\left(x_c^{\,1-\frac{hm}{n-1}}-x_0^{\,1-\frac{hm}{n-1}}\right) & \text{for } \dfrac{hm}{n-1}\neq 1, n\neq 1 \\ x_c\tan\theta-\left(\dfrac{v}{K}\right)^{\frac{1}{n-1}}\ln\left(\dfrac{x_c}{x_0}\right) & \text{for } \dfrac{hm}{n-1}=1, n\neq 1 \end{cases} \tag{4}$$

with x_c from equation (3). Plotting z_p against v gives a graphical solution for the divide velocity for a given total relief across the basin and set of physical parameters (Fig. 3c). A root-finding algorithm applied to equation (4) provides v for a given z_p. There are five independent parameters that go into this solution: θ, x_0, n, hm and v/K. In particular, K and v scale directly and so cannot be separated without independent information.

There is no non-trivial solution for $n=1$, which is troubling given that this is the value preferred by many studies. However, we can obtain a solution that considers the behaviour to the transient incision; this behaviour has been confirmed with numerical models. It is apparent that the fluvial channel at large x values will always incise more quickly than the advection can maintain relief, so the channel cuts down to the equilibrium gradient, thereby steepening and lengthening the hill slope above. Conversely, for small x values, advection of the steep hill slope will always outpace incision. With a simple threshold slope condition, there is no limit on the length of the hill slope, so the hill slope lengthens as the fluvial domain collapses to a single point at $x_c=z_c/\tan\theta$, and if slope continuity is maintained, the slope of the channel will be $tan\theta$ and the drainage area will scale with x_c, so we can calculate the divide velocity explicitly from equation (1) as:

$$v=K\left(\frac{z_p}{\tan\theta}\right)^{hm}$$

A diffusive hill slope would behave similarly in that the steepness of the hill slope can trade off against the advective velocity, so that the hill slope lengthens to take up all the relief, but velocity remains indeterminate as there are an infinite number of diffusive profiles to attain that relief. Very long, steep hill slopes grading to equilibrium channels are not commonly observed in nature, which either suggests that $n>1$ or points to a gap in our understanding of hillslope processes and their rate laws at high relief, as well as processes that are active at the transition between hillslope and channel processes.

The solution above can also be expressed in non-dimensional form by scaling by x_0:

$$\hat{z}_p=\frac{z_p}{x_0}$$

$$\hat{x}_c=\frac{x_c}{x_0}$$

$$\hat{x}_c=(\tan\theta)^{-\frac{n-1}{hm}}\left(\frac{v}{Kx_0^{hm}}\right)^{\frac{1}{hm}}$$

$$\hat{z}_p=\hat{x}_c\tan\theta-\frac{1}{1-\frac{hm}{n-1}}\left(\frac{v}{Kx_0^{hm}}\right)^{\frac{1}{n-1}}\left(\hat{x}_c^{\,1-\frac{hm}{n-1}}-1\right)$$

$$\frac{hm}{n-1}\neq 1, n\neq 1$$

This eliminates x_0 as a parameter, giving the non-dimensional group v/Kx_0^{hm}, the threshold slope and the exponents as parameters.

The velocity for $n=1$ also scales to:

$$\hat{z}_p=\tan\theta\left(\frac{v}{Kx_0^{hm}}\right)^{\frac{1}{hm}}$$

Although this solution is general for a given θ, hm and n, we find it more convenient to keep the problem dimensional, as it is otherwise difficult to see intuitively the parameter dependence (Fig. 3c).

Solution for a transport-limited river. A transport-limited river is one in which the sediment load in a river is always equal to its transport capacity, and the load adjusts by erosion or sedimentation. If the transport capacity is taken to be a function of slope and discharge it can be expressed as a power law in terms of drainage area: $Q=K_tS^nA^m$, where Q is the total sediment-carrying capacity of a river channel with upstream area A and channel slope S. 'Transport limitation' implies that incision or deposition is given by the divergence of sediment-transport capacity downstream, so that the mass balance can be expressed for a steady-state channel as:

$$\frac{\mathrm{d}z}{\mathrm{d}t}=0=-vhx^{h-1}\frac{\mathrm{d}z}{\mathrm{d}x}-K_t\frac{\mathrm{d}}{\mathrm{d}x}\left(x^{hm}\left(-\frac{\mathrm{d}z}{\mathrm{d}x}\right)^{n}\right) \tag{5}$$

As above, the negative sign on slope keeps the problem in the real domain, and we use Hack's Law to define the drainage area and include the basin width in the advective term. Basin width, W, is also defined from Hack's Law as:

$$W=\frac{\mathrm{d}A}{\mathrm{d}x}=hx^{h-1}$$

© 2018 Springer Nature Limited. All rights reserved.

Equation (5) can be integrated to give a solution in terms of the channel slope. This can be seen more clearly by reducing the differential equation to first order with the transformation:

$$s = \left(-\frac{\mathrm{d}z}{\mathrm{d}x}\right)^n$$

such that

$$\frac{vh}{K_t}(x^{-hm}s)^{\frac{1}{n}}x^{h-1-\frac{hm}{n}} = \frac{\mathrm{d}}{\mathrm{d}x}(x^{hm}s)$$

This can be factored, so as to be integrated, although for general exponents, there are multiple cases to be considered. In each case, we can solve for the slope to within a constant of integration, c.

$$\frac{vh}{K_t}x^{h-1-\frac{hm}{n}} = \frac{\mathrm{d}}{\mathrm{d}x}(x^{hm}s)\,(x^{hm}s)^{-\frac{1}{n}}$$

The integral is given for different cases as follows.
Case 1, where $m \neq n, n \neq 1$:

$$\frac{vx^{\frac{hn-hm}{n}}}{(n-m)K_t} = \frac{(x^{hm}s)^{\frac{n-1}{n}}}{n-1} + c$$

$$\frac{\mathrm{d}z}{\mathrm{d}x} = s^{\frac{1}{n}} = \left(\frac{(n-1)vx^{\frac{hn-hmn}{n}}}{(n-m)K_t} - (n-1)cx^{\frac{-hm(n-1)}{n}}\right)^{\frac{1}{n-1}} \quad (6)$$

Case 2, where $m \neq 1, n = 1$:

$$\frac{v}{(1-m)K_t}x^{h-hm} = \ln(x^{hm}s) - \ln c$$

$$\frac{\mathrm{d}z}{\mathrm{d}x} = -cx^{-hm}\exp\left(\frac{v}{(1-m)K_t}x^{h-hm}\right)$$

Case 3, where $m = 1, n = 1$:

$$\frac{vh}{K_t}\ln(x) = \ln(x^h s) - \ln c$$

$$\frac{\mathrm{d}z}{\mathrm{d}x} = -cx^h\exp\left(\frac{vh}{K_t}x\right)$$

Case 4, where $m = n, n \neq 1$:

$$\frac{v}{K_t}\ln(x) = \frac{(x^{hn}s)^{1-\frac{1}{n}}}{n-1} + c$$

$$\frac{\mathrm{d}z}{\mathrm{d}x} = x^{-h}\left((n-1)\left(\frac{v}{K_t}\ln(x) - c\right)\right)^{\frac{1}{n-1}}$$

Although the slope can be explicitly determined for any exponent, it is difficult to integrate these expressions to obtain $z(x)$ for a general exponent. However, this is not necessary in that we can determine boundary conditions for the slope and determine the advection velocity parametrically using these boundary conditions. At steady state, the total sediment generated upstream of a point x is $Q = v[z_p - z(x)]hx^{h-1}$. This can be equated to the capacity of the river at that point:

$$Q_c = K_t x^{hm}\left(-\frac{\mathrm{d}z}{\mathrm{d}x}\right)^n$$

Elevation, z, is not known in general, but it is known at the two points at the top and bottom of the river, x_c and x_0, where the elevation is z_c and 0, respectively. We can apply this condition at either point in order to determine the integration constant, c. The second point can then be used to determine v. We consider only the first two cases from above.

For case 1, where $m \neq n, n \neq 1$, we apply the flux condition at x_0. The slope solution is equation (6):

$$\frac{\mathrm{d}z}{\mathrm{d}x} = s^{\frac{1}{n}} = \left(\frac{(n-1)vx^{\frac{hn-hmn}{n}}}{(n-m)K_t} - (n-1)cx^{\frac{-hm(n-1)}{n}}\right)^{\frac{1}{n-1}}$$

subject to:

$$K_t x_0^{hm}s = vz_p h x_0^{h-1}$$

giving c as:

$$c = \frac{vx_0^{\frac{hn-hm}{n}}}{(n-m)K_t} - \frac{1}{n-1}\left(\frac{vz_p h}{K_t}\right)^{\frac{n-1}{n}}x_0^{\frac{(h-1)(n-1)}{n}}$$

With the threshold slope condition as above, we can solve for x_c:

$$K_t x_c^{hm}(\tan\theta)^n = vx_c\tan\theta h x_c^{h-1}$$

$$x_c = \left(\frac{vh}{K_t}\right)^{\frac{1}{h(m-1)}}(\tan\theta)^{\frac{1-n}{h(m-1)}}, m \neq 1$$

Equation (6), using x_c and the threshold slope, is then defined as:

$$\tan\theta = \left(\frac{(n-1)vx_c^{\frac{hn-hmn}{n}}}{(n-m)K_t} - (n-1)cx_c^{\frac{-hm(n-1)}{n}}\right)^{\frac{1}{n-1}}$$

with x_c and c as above. This can be solved for z_p to give a parametric relationship between velocity and z_p:

$$z_p = \frac{K_t}{vh}\left[(\tan\theta)^{n-1}x_c^{\frac{hm(n-1)}{n}}x_0^{\frac{-hn+h+n-1}{n}} + \frac{(n-1)v}{(n-m)K_t}\left(x_0^{\frac{h+n-1-hm}{n}} - x_c^{\frac{hn-hm}{n}}x_0^{\frac{-hn+h+n-1}{n}}\right)\right]^{\frac{n}{n-1}}$$

For case 2, where $m \neq 1, n = 1$:

$$\frac{v}{(1-m)K_t}x^{h-hm} = \ln(x^{hm}s) - \ln c$$

We apply flux condition at x_0, requiring that:

$$K_t x_0^{hm}s = vz_p h x_0^{h-1}$$

giving c as:

$$c = \left(\frac{vz_p h x_0^{h-1}}{K_t}\right)\exp\left(-\frac{v}{(1-m)K_t}x_0^{h-hm}\right)$$

We can assume continuity of slope at the transition, x_c, obtaining the condition that can be used to solve for x_c:

$$K_t x_c^{hm}(\tan\theta) = vx_c\tan\theta h x_c^{h-1}$$

$$x_c = \left(\frac{vh}{K_t}\right)^{\frac{1}{h(m-1)}}, m \neq 1$$

The slope equation with x_c and c as above can be solved for z_p to give a parametric relationship between velocity and z_p:

$$z_p = \frac{K_t\tan\theta}{vh}x_0^{1-h}x_c^{hm}\exp\left(-\frac{v}{(1-m)K_t}x_c^{h-hm}\right)$$

Data availability

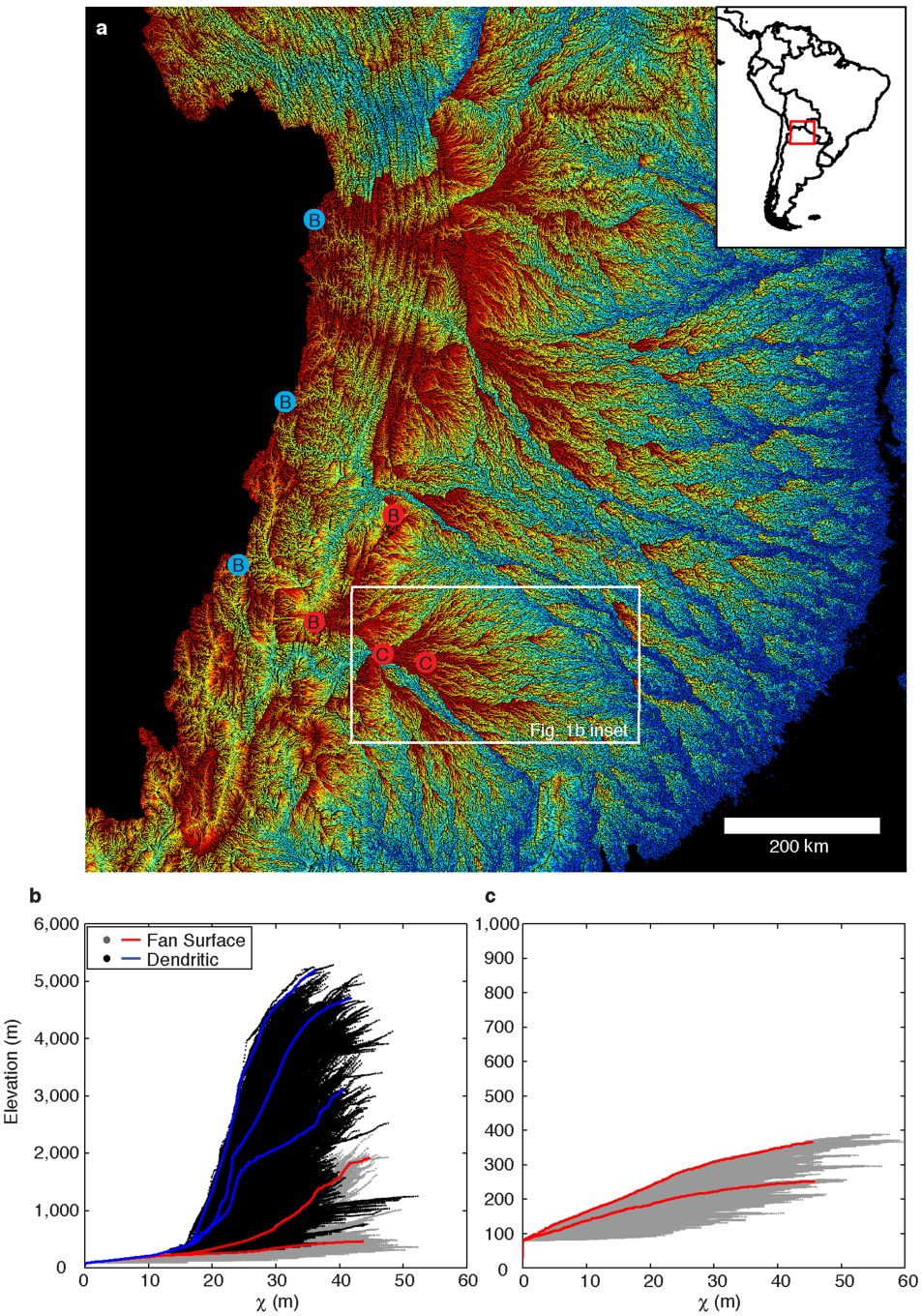
Topographic data used in support of this study are publicly available through SRTM and downloadable from Open Topography (http://www.opentopography.org/). The locations of playa lakes are available from ref. [32].

34. National Wetlands Inventory. *US Department of the Interior, Fish and Wildlife Service* http://www.fws.gov/wetlands/ (2017).

© 2018 Springer Nature Limited. All rights reserved.

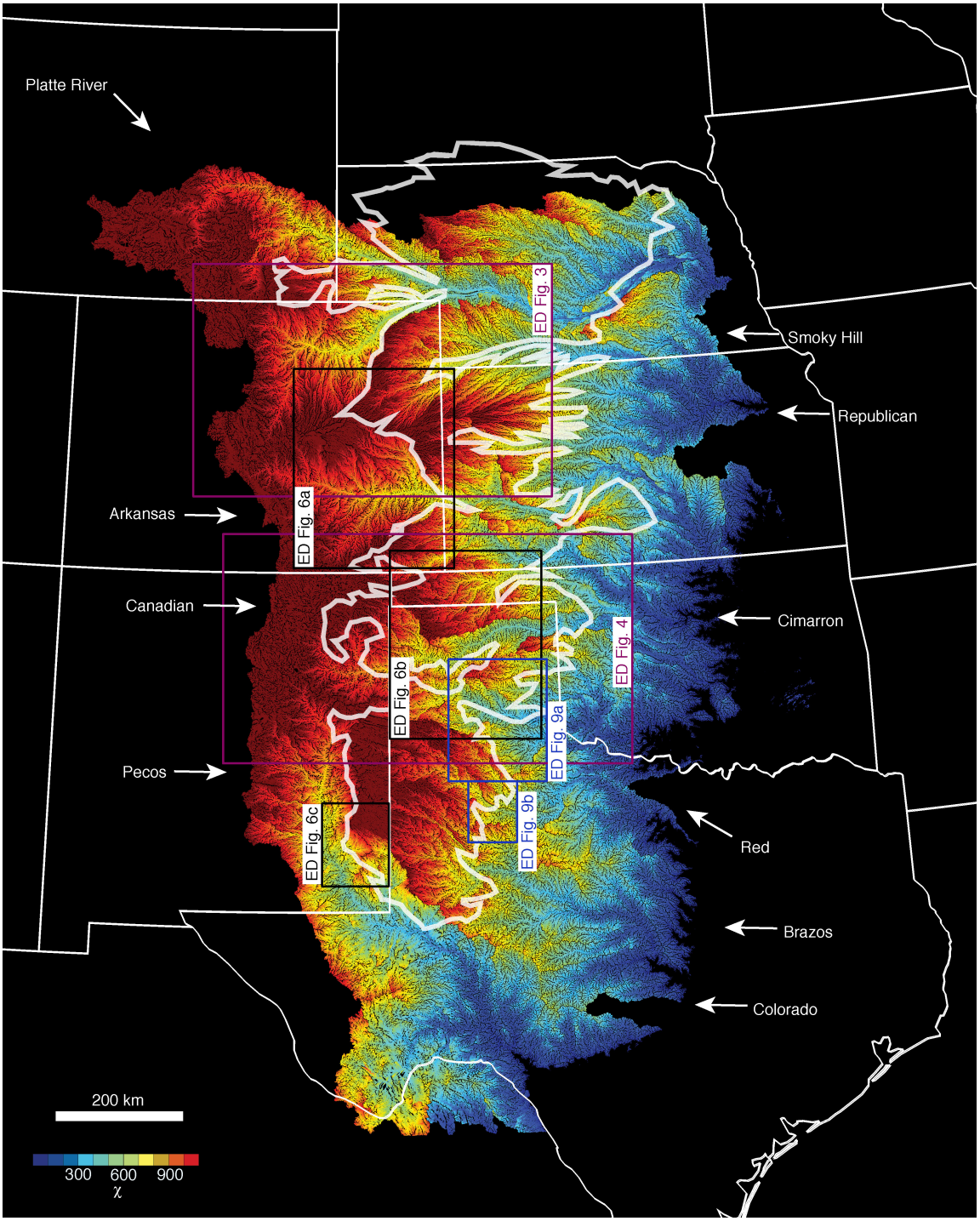
35. Homer, C. G. et al. Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information. *Photogramm. Eng. Remote Sensing* **81**, 345–354 (2015).
36. Holliday, V. T. Soils and landscape evolution of eolian plains: the Southern High Plains of Texas and New Mexico. *Geomorphology* **3**, 489–515 (1990).
37. Braun, J. A review of numerical modeling studies of passive margin escarpments leading to a new analytical expression for the rate of escarpment migration velocity. *Gondwana Res.* **3**, 209–224 (2018).
38. Perron, J. T. & Hamon, J. L. Equilibrium form of horizontally retreating, soil-mantled hillslopes: model development and application to a groundwater sapping landscape. *J. Geophys. Res. Earth Surf.* **117**, F01027 (2012).

© 2018 Springer Nature Limited. All rights reserved.



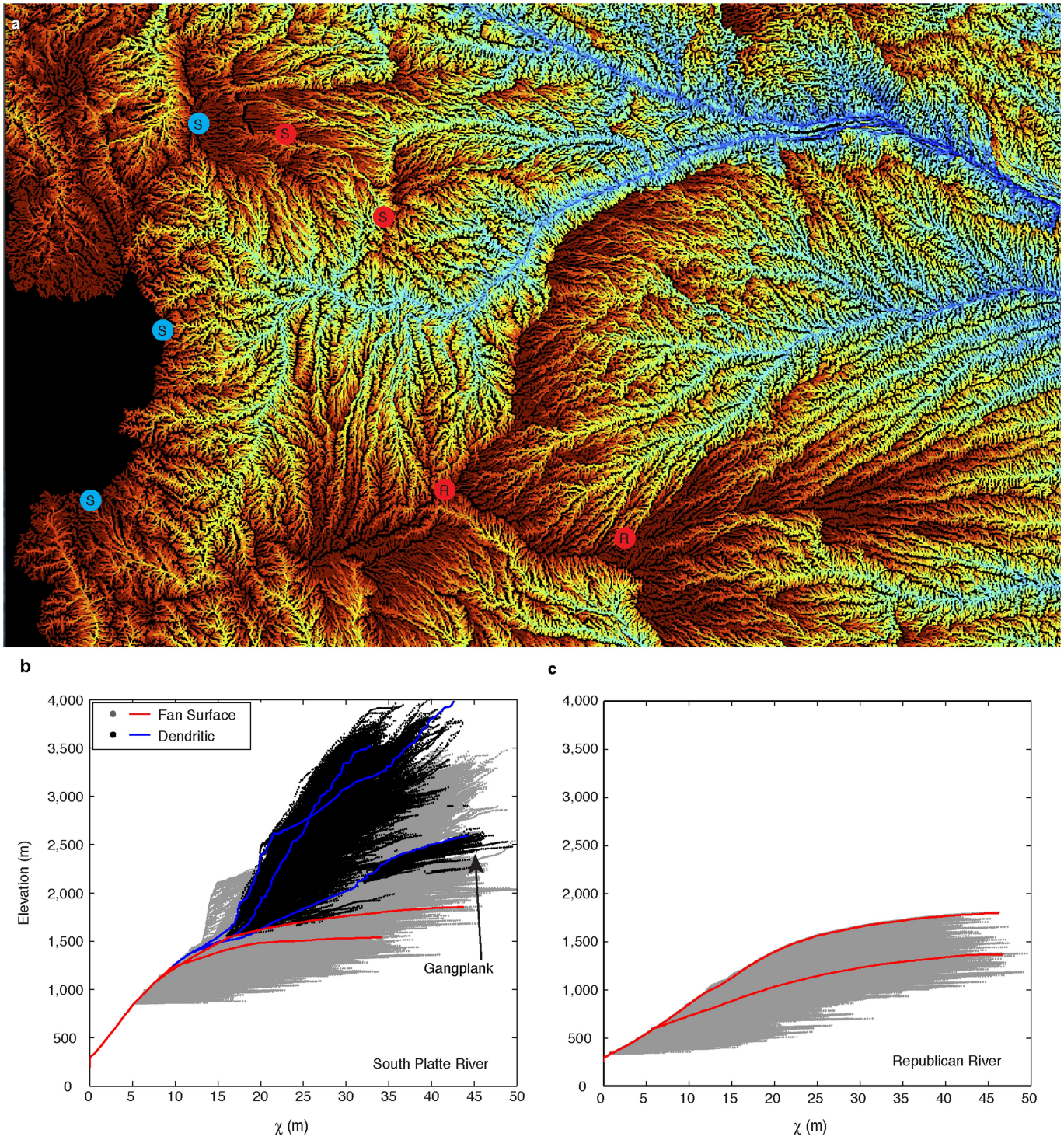
Extended Data Fig. 1 | χ map and χ versus elevation plots for modern-day South American megafans, constructed using a concavity of 0.45. **a**, χ map, showing megafans formed in the Andean foreland of Bolivia and Argentina, with scaling as in Fig. 1. The Plicomayo River forms the northernmost prominent fan on the border of Bolivia and Argentina. The Juramento River forms the southernmost prominent fan. The Bermejo River forms the central fan. Points indicate channel heads of rivers highlighted in lower panels: B, channel heads for profiles shown in panel **b**; C, channel heads for profiles shown in panel **c**. Inset, map of South America, with the extent of the χ map outlined in red. **b**, Plot of χ versus elevation for the Plicomayo, Bermejo and Juramento river basins. Full sets of dendritic and fan-draining rivers define the black and grey envelopes, respectively; blue and red profiles show typical rivers from within each data cloud. **c**, Plot of χ versus elevation for the currently inactive portion of the fan.

© 2018 Springer Nature Limited. All rights reserved.



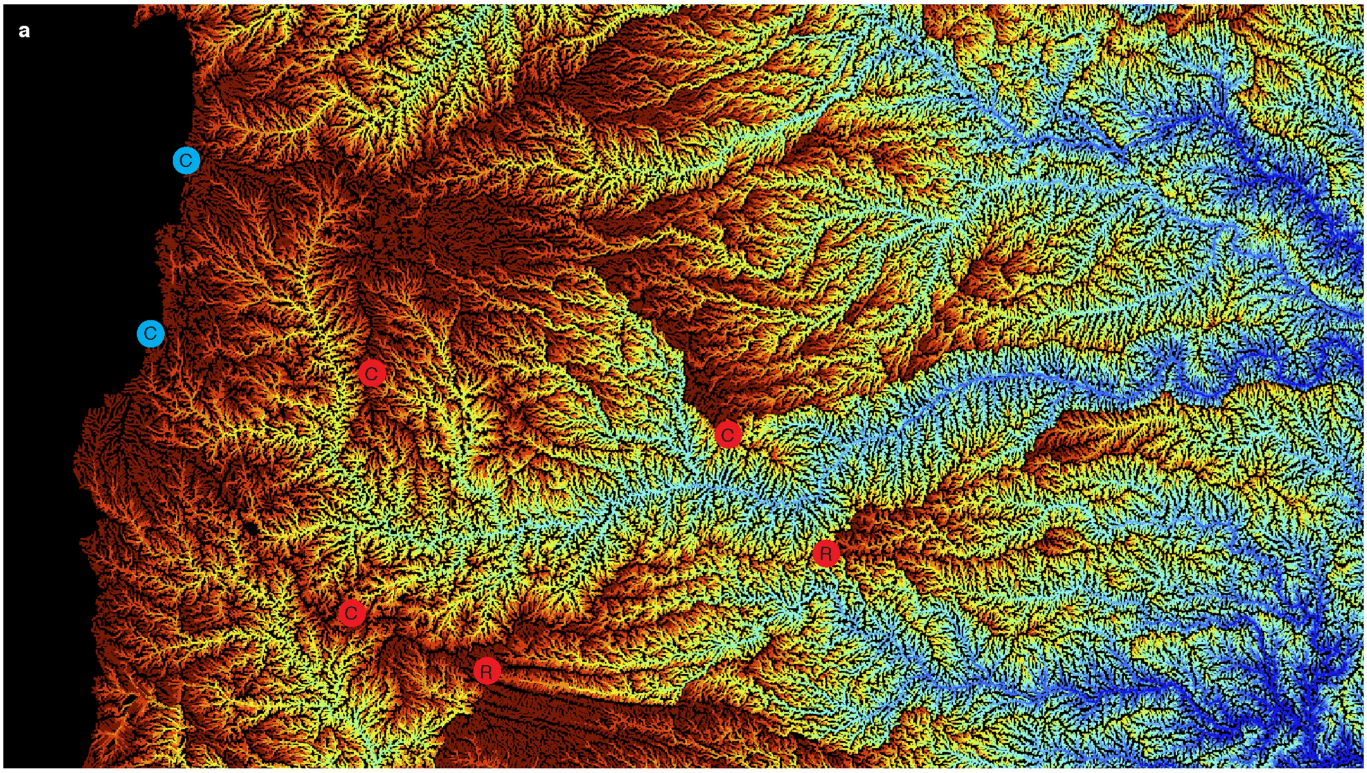
Extended Data Fig. 2 | Map of the river network in the Rocky Mountains and High Plains delineated by χ, calculated using a concavity of 0.3. Patterns and major features do not differ greatly from those in the map constructed using a concavity of 0.45 (Fig. 1). The thick white line shows the extent of the Ogallala Group sediments. Thin white lines mark state boundaries. Labels denote river basins mentioned in the main text. River basins draining the Rocky Mountains are labelled near their headwaters; river basins draining fan surfaces covered in Ogallala sediment are labelled near their outlets. Coloured boxes show the locations of data provided in the indicated figures.

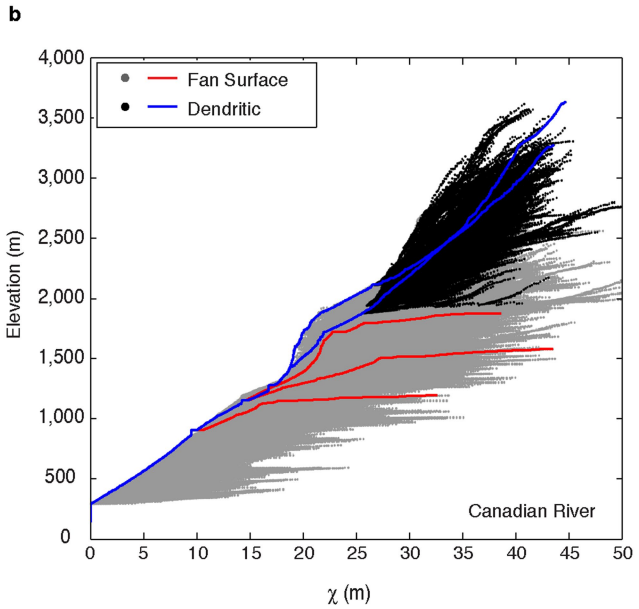
© 2018 Springer Nature Limited. All rights reserved.

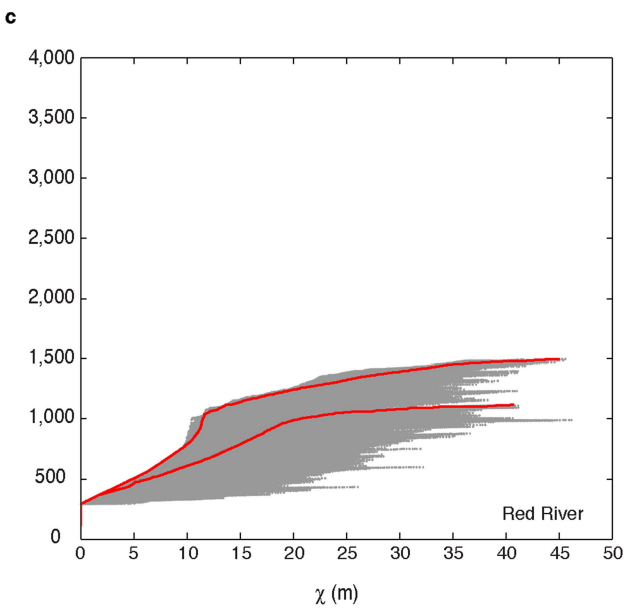


Extended Data Fig. 3 | χ maps and χ versus elevation plots for the South Platte and Republican river basins. a, χ map centred on the South Platte river basin, with the same colour scale as Fig. 1 (location shown in Extended Data Fig. 2). Points indicate the channel heads of rivers highlighted in the lower panels: S, South Platte River; R, Republican River. **b**, Plot of χ versus elevation for the South Platte river basin. Full sets of dendritic and fan-draining rivers define the black and grey envelopes, respectively; blue and red profiles show typical rivers from within each data cloud. The lower steepness tributary marked as the Gangplank drains a fan remnant that marks the farthest western extent of the Ogallala Group. **c**, χ/elevation plot for the Republican river basin. The main stems of dendritic rivers that largely drain the Rocky Mountains (blue lines) have large drainage areas and near-linear χ/elevation plots characteristic of an integrated river that is close to equilibrium. By contrast, rivers that drain surfaces capped by the Ogallala Group fan sediments show curved, divergent channel profiles (red lines), consistent with single-channel, non-branching rivers flowing directly down a planar surface.

© 2018 Springer Nature Limited. All rights reserved.

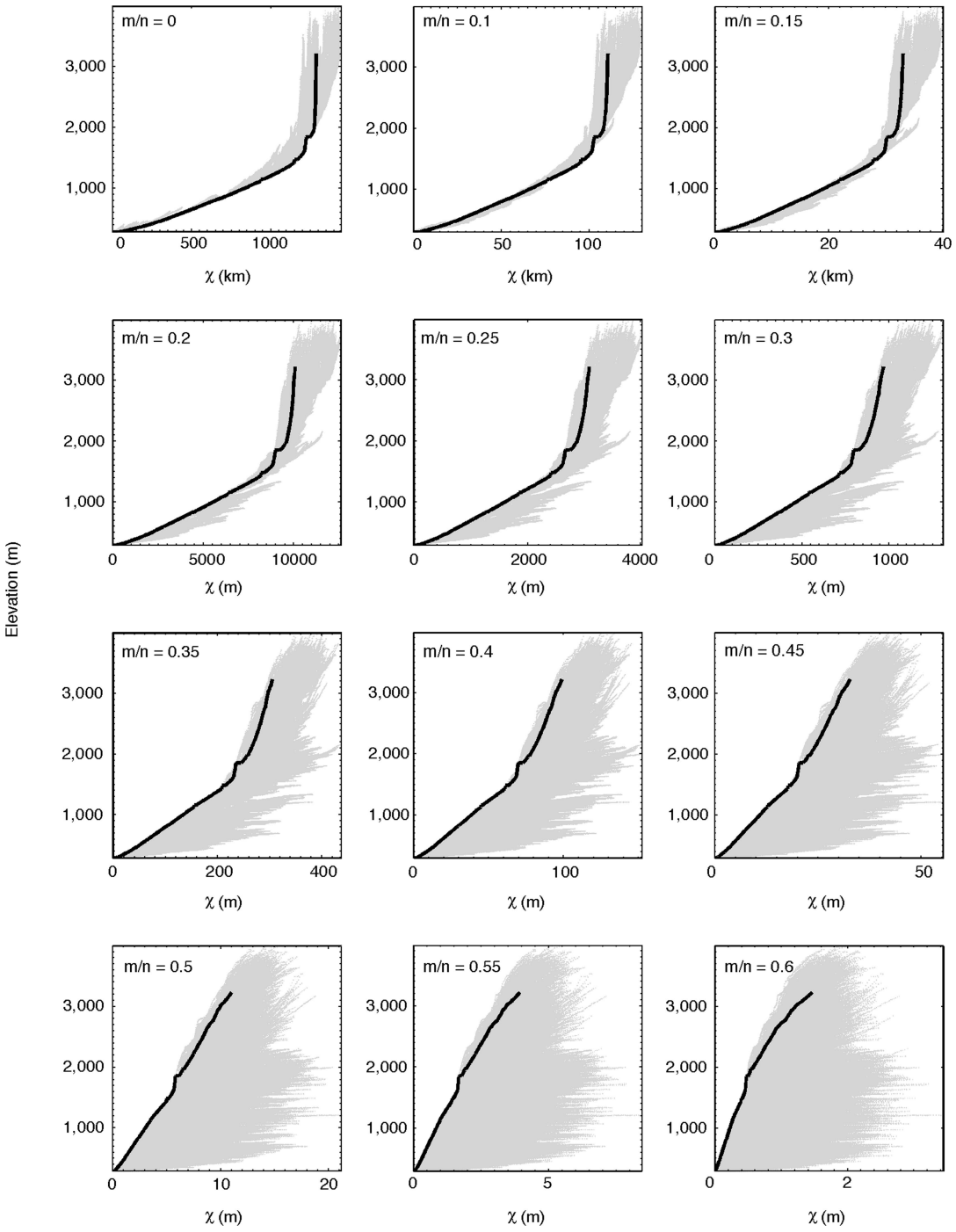






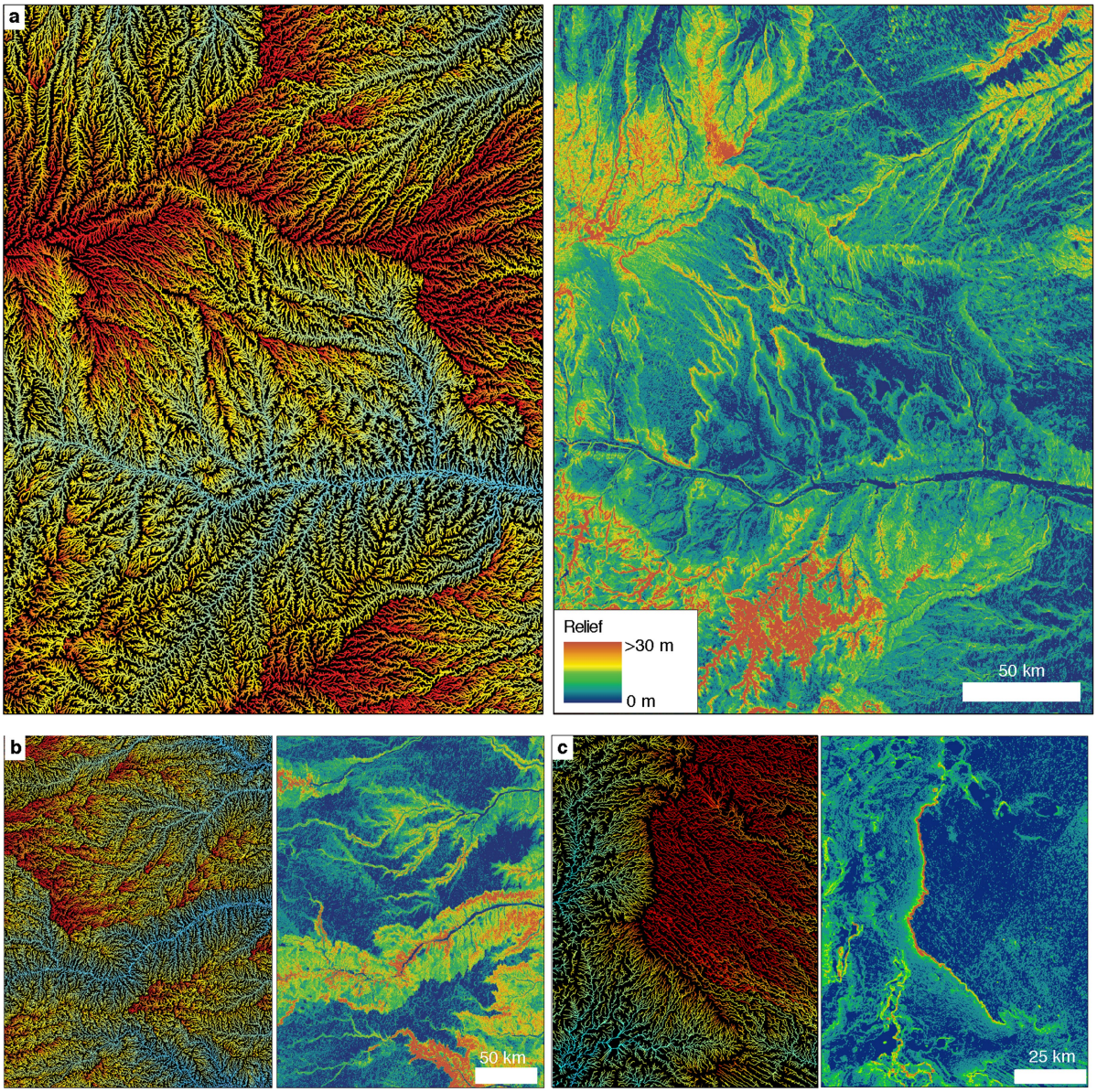
Extended Data Fig. 4 | Scaled distance maps and elevation profiles for the Canadian and Red river basins. a, χ map centred on the Canadian river basin, with the same colour scale as Fig. 1 (location shown in Extended Data Fig. 2). Points indicate channel heads of rivers highlighted in the lower panels: C, Canadian River; R, Red River. **b**, χ-elevation plot for the Canadian river basin. **c**, χ-elevation plot for the Red river basin.

© 2018 Springer Nature Limited. All rights reserved.



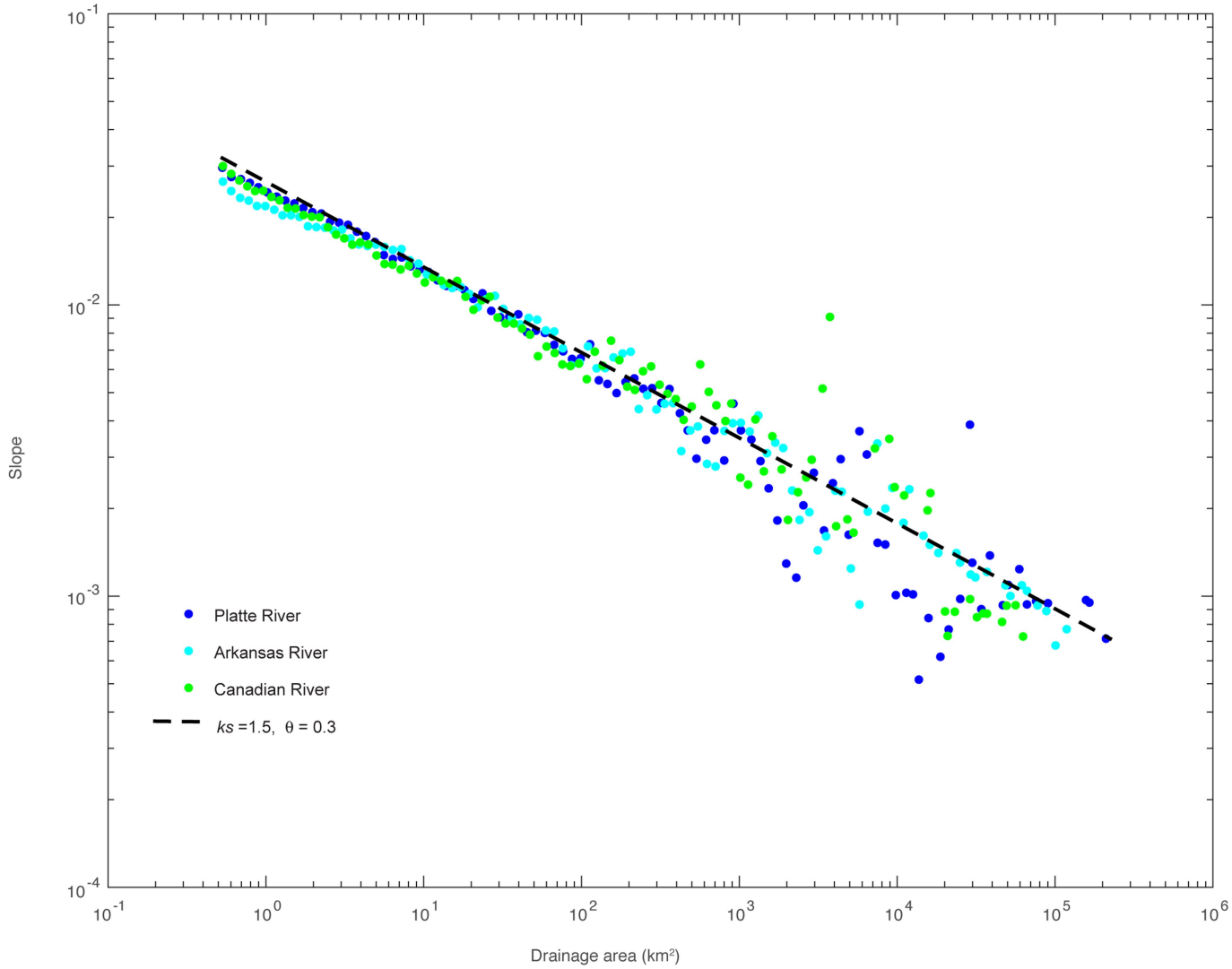
Extended Data Fig. 5 | Plots of χ versus elevation for the Arkansas river basin. Plots were constructed using a range of concavity values, $\theta = m/n$, from 0 to 0.6. The main stem of the Arkansas River is delineated with a solid black line. A concavity of 0.0 reduces χ to river distance from base level, so the upper left plot shows the longitudinal profiles of the river. We expect the main stems of large rivers draining the Rocky Mountains to be near equilibrium owing to their large drainage area and short response times. Slope/area analysis of the main stems of three of the major rivers (the Platte, Arkansas and Canadian rivers; Extended Data Fig. 7) indicates a concavity of about 0.3. With this concavity, the Arkansas River exhibits a concave-up, or kinked, form that is consistent with the harder, less erosive rock in the Rocky Mountains, and/or with the eastward tilt of the region.

© 2018 Springer Nature Limited. All rights reserved.



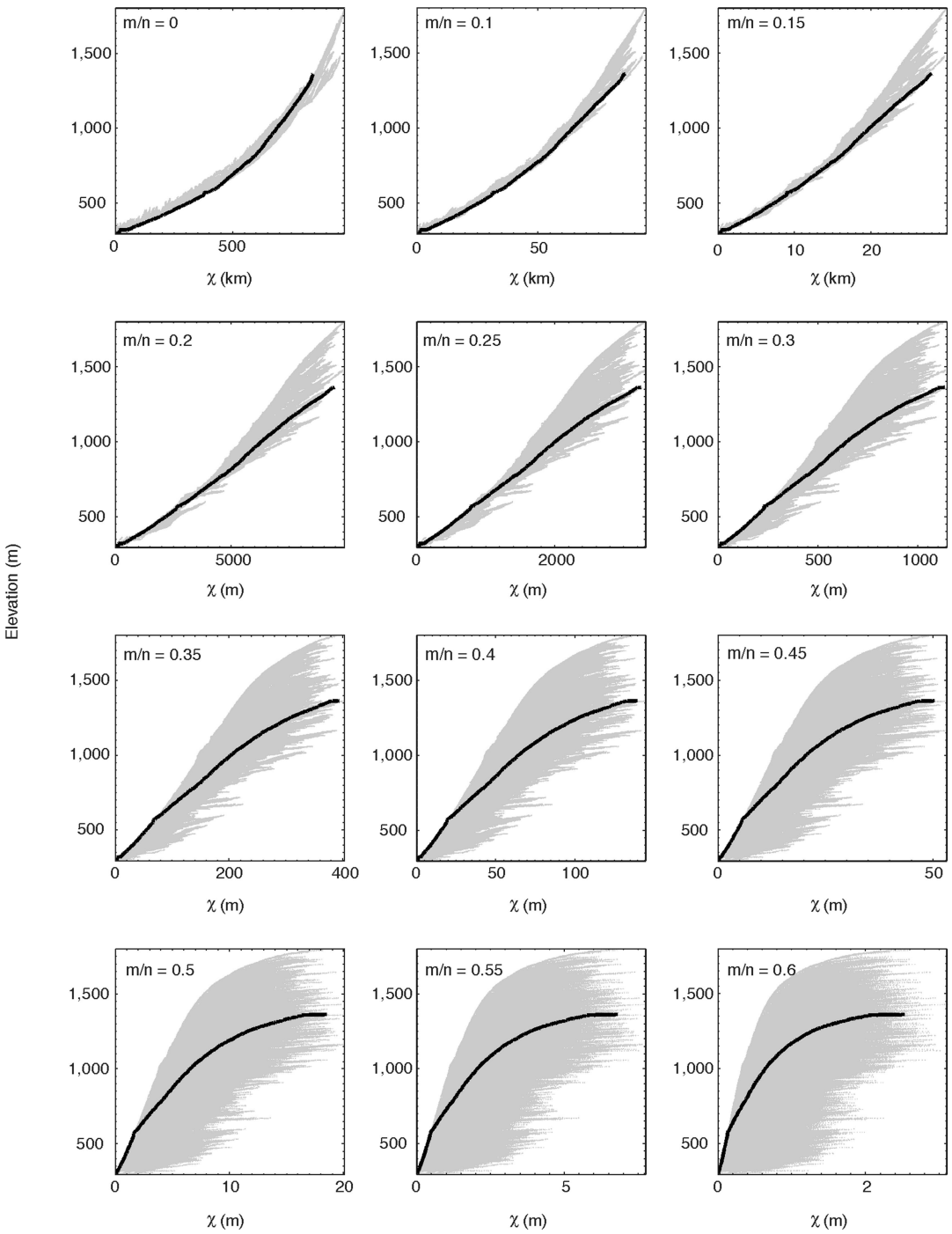
Extended Data Fig. 6 | Paired χ and relief maps. These maps are centred on: **a**, the divide separating the Arkansas river basin from the Smoky Hill river basin; **b**, the divides separating the lower Canadian river basin from the North Canadian river basin in the north, and from the Red river basin in the south; and **c**, the divide separating the upper Colorado river basin and the upper Pecos river basin. See Extended Data Fig. 2 for names and locations. All three examples show that low relief areas occur where the fluvial network is poorly integrated, and that large changes in relief correspond to large cross-divide χ gradients. An exception is visible in panel **b**, centre left, where a large cross-divide χ gradient has no corresponding relief. This corresponds to a large river capture that transfers a tributary of the North Canadian to the lower Canadian River.

© 2018 Springer Nature Limited. All rights reserved.



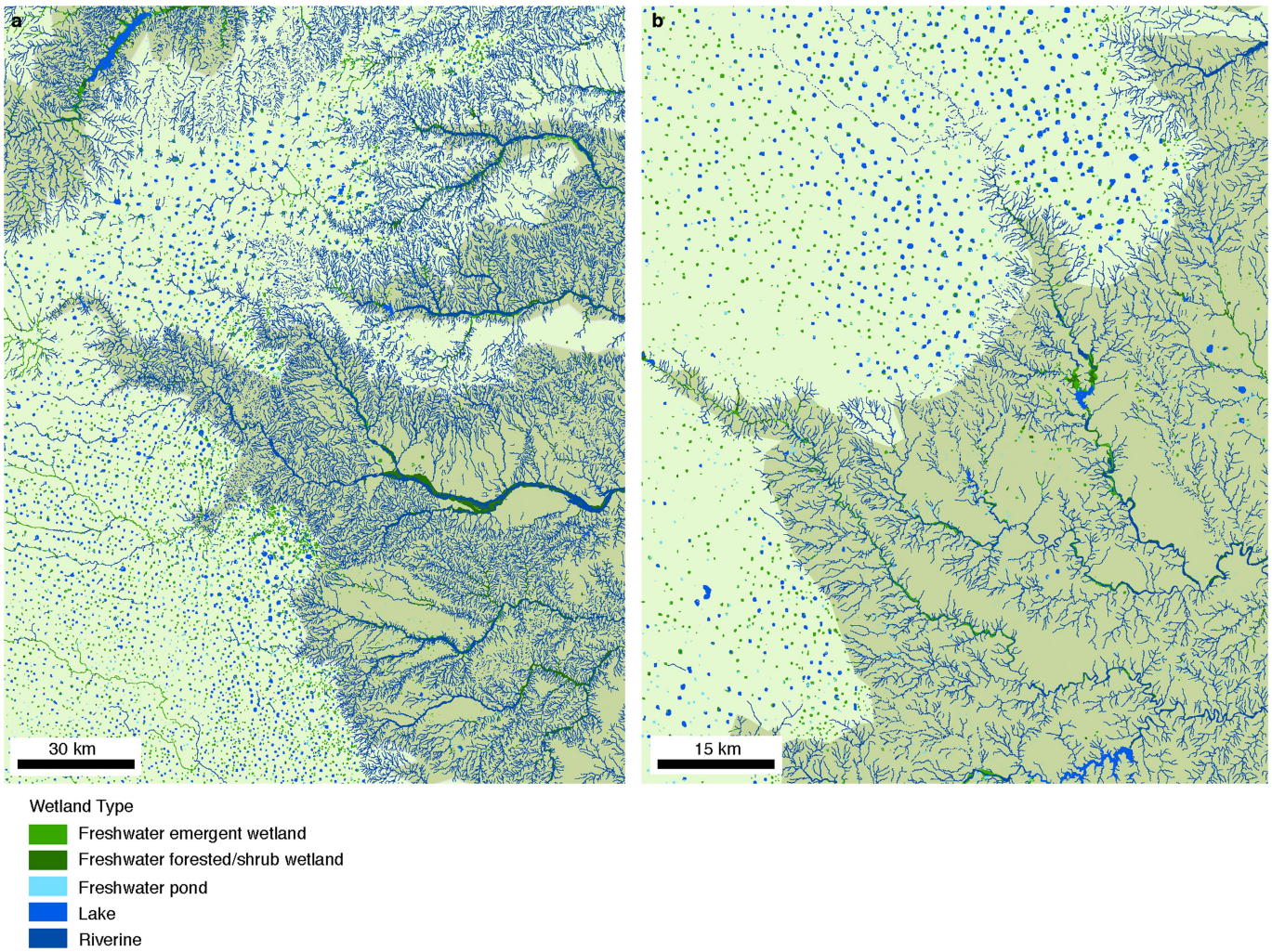
Extended Data Fig. 7 | Slope/area regression in log space for the Platte, Arkansas and Canadian rivers. Simultaneous regression for all three rivers yields a concavity, θ, of 0.3 and a steepness, k_s, of 1.5.

© 2018 Springer Nature Limited. All rights reserved.



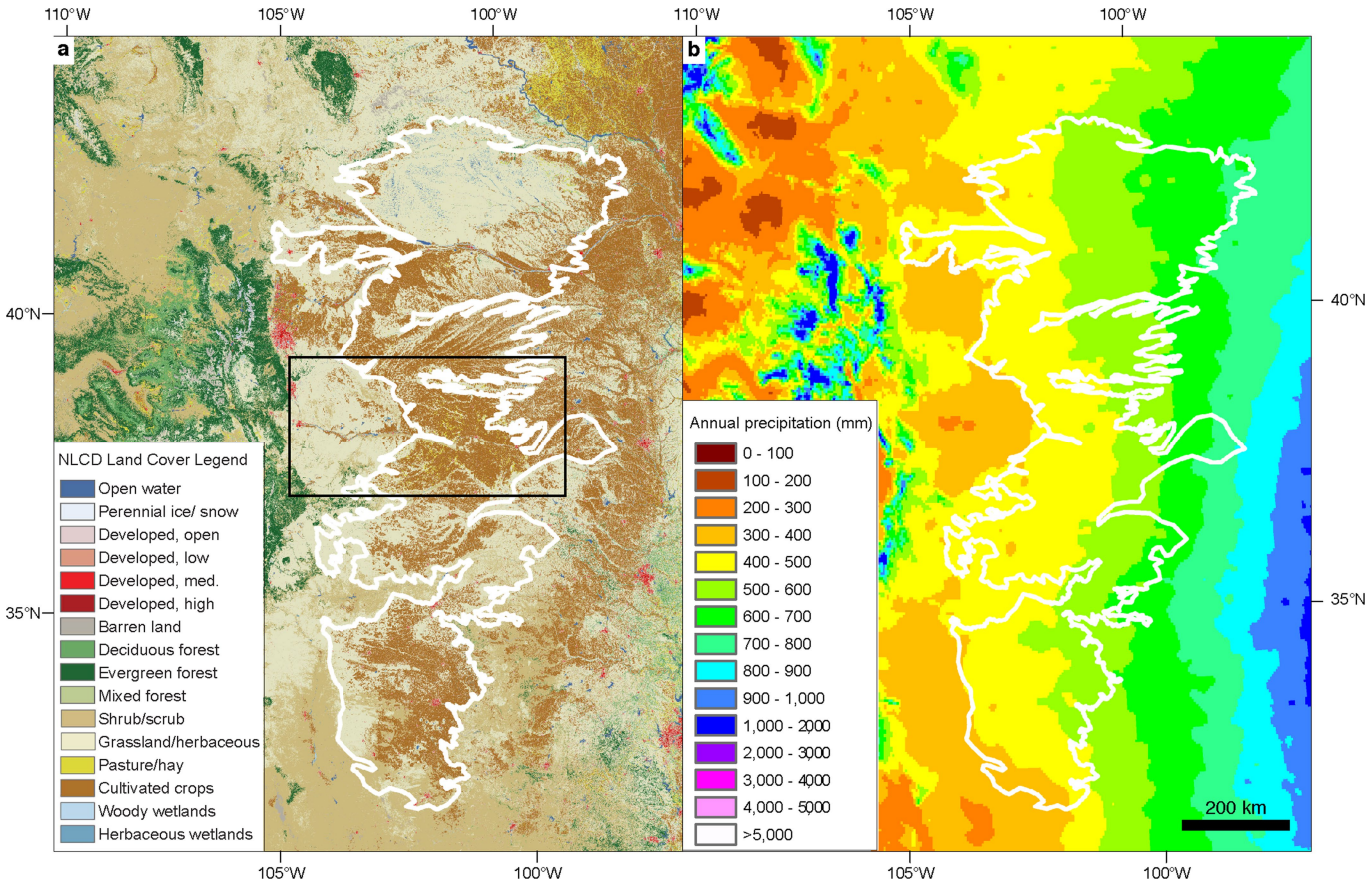
Extended Data Fig. 8 | Plots of χ versus elevation for the Republican river basin. Plots were constructed using a range of concavity values, $\theta = m/n$, from 0 to 0.6. The main stem of the Republican River is delineated with a solid black line. The Republican river basin primarily drains fans that are capped by sediments of the Ogallala Group. The collapse of χ plots for $m/n \approx 0.0$ and the curved, divergent channel profiles for $m/n > 0.0$ highlight that these rivers are single-channel, almost non-branching rivers flowing down an inclined planar surface.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Maps showing the distribution of wetland types on and off the Ogallala surface. a, **b**, Maps centred on the eastern escarpments of the Red (**a**) and Brazos (**b**) Rivers. See Extended Data Fig. 2 for names and locations; see Methods for further details. The Ogallala surface is shown in light green and is covered in isolated wetlands (lakes and freshwater ponds). Away from the Ogallala surface, isolated wetlands are scarce: wetlands exist only along the river network, and are predominantly of the riverine and freshwater emergent and forested/shrub type.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Maps showing land cover and mean annual precipitation for the Ogallala Group. a, Land cover. **b**, Mean annual precipitation, based on 30-year normal data for 1981 to 2010, downloaded from the PRISM Climate Group (http://prism.oregonstate.edu/normals/). The white outline delineates the Ogallala Group sediments. Contrasts in land cover are apparent between areas from which the Ogallala sediments have been eroded (now largely grassland) compared with where they remain (higher density of cultivated crops). For areas in and around the Arkansas river basin (black rectangle), 14% of land cover comprises cultivated crops, where the Ogallala sediments have been eroded; 58% comprises cultivated crops, where the Ogallala sediments are preserved (see Methods). A similar pattern of land use is present throughout the High Plains region. For the region east of and proximal to the Rocky Mountain front, and south of the North Platte river basin (where eolian deposits prevent cultivation), 12% of land cover comprises cultivated crops, where the Ogallala sediments have been removed; for 42% of land cover, the Ogallala sediments remain (Methods).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0521-4

The role of miniaturization in the evolution of the mammalian jaw and middle ear

Stephan Lautenschlager[1,2]*, Pamela G. Gill[1,3], Zhe-Xi Luo[4], Michael J. Fagan[5] & Emily J. Rayfield[1]*

The evolution of the mammalian jaw is one of the most important innovations in vertebrate history, and underpins the exceptional radiation and diversification of mammals over the last 220 million years[1,2]. In particular, the transformation of the mandible into a single tooth-bearing bone and the emergence of a novel jaw joint—while incorporating some of the ancestral jaw bones into the mammalian middle ear—is often cited as a classic example of the repurposing of morphological structures[3,4]. Although it is remarkably well-documented in the fossil record, the evolution of the mammalian jaw still poses the paradox of how the bones of the ancestral jaw joint could function both as a joint hinge for powerful load-bearing mastication and as a mandibular middle ear that was delicate enough for hearing. Here we use digital reconstructions, computational modelling and biomechanical analyses to demonstrate that the miniaturization of the early mammalian jaw was the primary driver for the transformation of the jaw joint. We show that there is no evidence for a concurrent reduction in jaw-joint stress and increase in bite force in key non-mammaliaform taxa in the cynodont–mammaliaform transition, as previously thought[5–8]. Although a shift in the recruitment of the jaw musculature occurred during the evolution of modern mammals, the optimization of mandibular function to increase bite force while reducing joint loads did not occur until after the emergence of the neomorphic mammalian jaw joint. This suggests that miniaturization provided a selective regime for the evolution of the mammalian jaw joint, followed by the integration of the postdentary bones into the mammalian middle ear.

The mammalian jaw and jaw joint are unique among vertebrates[6]. The craniomandibular jaw joint of non-mammalian vertebrates is formed between the quadrate and articular bones, whereas mammals evolved a novel jaw hinge between the squamosal and dentary bones—the secondary or temporomandibular jaw joint[1–4]. The evolutionary origins of this morphological transformation involved a suite of osteological modifications to the feeding and auditory systems that occurred over a period of 100 million years during the Late Triassic and Jurassic across the cynodont–mammaliaform transition[9,10]. The tooth-bearing dentary bone increased in size relative to the postdentary elements, eventually transforming the seven-bone lower jaw in pre-mammalian cynodonts (hereafter referred to as cynodonts) to a single-bone lower jaw in modern mammals; parallel to this simplification of the mandible, the integration of elements of the ancestral craniomandibular jaw joint into the ossicular chain led to a unique middle and inner ear morphology capable of more sensitive sound detection[11,12]. New fossil information has suggested that a definitive mammalian middle ear (DMME) evolved independently in at least three mammalian lineages by detachment from the mandible, but the emergence of a secondary jaw joint is a key innovation that unites all mammaliaforms[9,13]. However, a central question exists as to how, during this transformation, the jaw hinge remained robust enough to bear strong mastication forces while the same bones were becoming delicate enough to be biomechanically viable for hearing[3,5,10].

The stepwise acquisition of morphological features leading to the emergence of the temporomandibular jaw joint is exceptionally well-documented in the fossil record, by a series of transitional taxa that illuminates the evolutionary dynamics involved[4]. While still appearing to function as a jaw joint and viable for sound transmission in cynodonts (for example, *Thrinaxodon liorhinus*, *Probainognathus* and *Probelesodon sanjuanensis*), the postdentary bones gradually reduced in size and shifted away from the jaw joint—probably for more sensitive hearing[10,12]. This trend resulted in all basal mammaliaforms (for example, *Sinoconodon rigneyi* and *Morganucodon oehleri*) possessing a remarkable 'dual jaw joint' with two seemingly functional joints: a quadrate–articular joint medial to a mammalian dentary condyle and squamosal glenoid hinge[11,13]. More derived groups and crown mammals eventually lost the ancestral quadrate–articular joint. In addition to fossil evidence, this sequence of events was identified historically in embryonic stages of living mammals[14,15] and recent morphogenetic studies, gene patterning and regulatory networks have further elucidated the development of these structures[16,17]. Previous studies have theorized that muscle reorganization reduced load at the jaw joint[6,10], yet these claims have not been tested in fossil taxa and experimental studies of extant mammals reveal that the jaw joint usually experiences net compressive loading[18,19]. The modification of the mandible and the emergence of a novel jaw joint and middle ear therefore represent an intriguing problem. This is especially puzzling when all the evidence points towards modifications for increased jaw muscle force, consolidation of cranial bones, increased complexity of sutures and supposedly stronger skulls during mammalian evolution[1,5].

Here we integrate a suite of digital reconstruction, visualization and quantitative biomechanical modelling techniques to test the hypothesis that reorganization of the adductor musculature and reduced stress susceptibility in the ancestral jaw joint facilitated the emergence of the mammalian temporomandibular jaw joint. Applying finite element analysis, we calculated bone stress, strain and deformation to determine the biomechanical behaviour of the mandibles of six key taxa across the cynodont–mammaliaform transition (Fig. 1). These analyses were supplemented by multibody dynamics analysis to predict bite forces and joint reaction forces. Results from the combined analyses demonstrate that during simulated biting there is no evidence for the reduction of stresses (von Mises, tensile and compressive) in the jaw joint (craniomandibular jaw joint and/or temporomandibular jaw joint) across the cynodont and mammaliaform taxa that we studied (Figs. 2, 3, Extended Data Figs. 1, 3). This was found to be the case for unilateral and bilateral biting simulations, regardless of the working and balancing side joint. However, bite position appears to have a moderate effect on joint stresses—particularly compression—and stress increases as the bite point moves anteriorly along the tooth row. This is consistent with experimental data for extant mammals, in which incisor biting results in the highest joint loads[20]. Similarly, the results of multibody dynamics analysis show that absolute joint reaction forces are not reduced as the jaw joint undergoes morphological transformation (Figs. 2, 3), whereas relative bite forces—the ratio between muscle force and bite

[1]School of Earth Sciences, University of Bristol, Bristol, UK. [2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK. [3]Earth Science Department, The Natural History Museum, London, UK. [4]Department of Organismal Biology and Anatomy, University of Chicago, Chicago, IL, USA. [5]School of Engineering and Computer Science, University of Hull, Hull, UK. *e-mail: s.lautenschlager@bham.ac.uk; e.rayfield@bristol.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

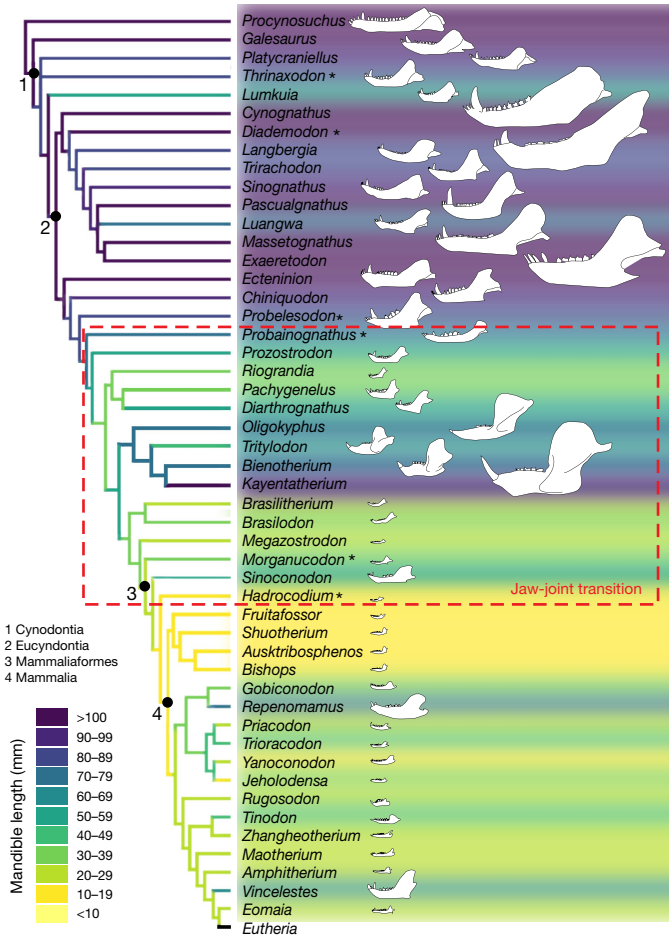


Fig. 1 | Mandibular sizes and evolutionary relationships of cynodonts, mammaliaforms and mammals. Asterisks denote taxa included in this study. The phylogeny is simplified after previously published phylogenies[13,28,29].

force—decrease in derived cynodonts (*Probainognathus* and crownward animals) and stay largely constant in mammaliaforms such as *M. oehleri* and *Hadrocodium wui* (Extended Data Figs. 1–3).

However, the simulation of different muscle activation patterns using finite element analysis reveals that, across the cynodont–mammaliaform transition, there is a distinct shift in the recruitment of jaw adductor musculature required to achieve high bite forces that maintain low stress in the jaw joint (Fig. 4a). In the cynodonts *T. liorhinus*, *Diademodon tetragonus* and *Probainognathus* sp., the highest bite forces that keep joint tensile stresses low are found when jaw adduction is dominated by the masseter muscle group. In *P. sanjuanensis* and the mammaliaforms *M. oehleri* and *H. wui*, the recruitment of the pterygoideus muscle group (with contributions from the masseter musculature) provides the highest relative bite forces, eventually shifting to the temporalis group as the dominant contributor for high bite force–low joint stress performance in the extant taxon *Monodelphis domestica*. This pattern is reversed for muscle activations that are optimized for high bite force–low compressive stresses in the jaw joint in mammaliaforms (Fig. 4a). This is achieved mainly by recruitment of the temporalis group in cynodonts, whereas the masseter and the pterygoideus groups form the dominant musculature in mammaliaforms. Apart from an overall shift in the pattern of muscle recruitment, the analyses further demonstrate that whereas in the cynodonts and mammaliaforms a single muscle group is harnessed to achieve 'optimal' bite forces (that is, high bite force–low jaw-joint stress), *M. domestica* simultaneously activates all three muscle groups (Extended Data Figs. 4–9) as revealed by our computational analyses, which confirm previous experimental data[21]. Changes in muscle orientation and inferred muscle lines of action either precede or are associated with mandible shape change[22], leading to a more efficient use of the adductor system to maximize bite force and minimize loads on the jaw joint.

Considering that the shape of the mandible alone does not appear to have a substantial influence on stress reduction in the mandibles of the taxa that we studied, we further tested size-related effects on the biomechanical behaviour of the jaw joint. All taxa were scaled to seven different jaw lengths (5, 10, 20, 40, 80, 160 and 320 mm) that cover the mandibular size range observed across the cynodont–mammaliaform transition (Figs. 1, 4b, Supplementary Table 1). Results for these hypothetical resized models demonstrate that absolute tensile and compressive stresses in the jaw joint decrease exponentially to 25% with a reduction in size by 50%. At the same time, absolute bite forces decrease by 50%, in direct proportion to jaw length.

We offer biomechanical evidence that the stress susceptibility of the mandible, and in particular of the jaw joint, was not reduced across the cynodont–mammaliaform transition. This contrasts with existing hypotheses that argue for a reduction of joint loads owing to the rearrangement of the jaw adductor musculature and the resultant increase in bite force in mammaliaforms[1,7,8]. The decrease in size of the postdentary bones and the accompanying expansion of bony angular and coronoid projections of the dentary is assumed to have paralleled the reorganization and evolution of mammalian masseter and temporalis muscle groups[6,8,21]. Arranged in such a manner, the changed line of action of the major jaw adductor muscles was hypothesized to have led to a redistribution of muscle forces with little or no load experienced at the jaw hinge. However, our results do not support these previous inferences.

As demonstrated here, a change in the recruitment of the jaw adductor musculature can be observed to achieve high bite forces and at the same time to keep tensile and compressive joint stresses at a minimum. These findings parallel experimental data from extant mammals that show that differential muscle activation produces different stress regimes in the jaw joint[20]. However, in the cynodonts and mammaliaforms that we studied, parallel activation of all three adductor muscle groups does not lead to the highest relative bite forces (Extended Data Figs. 4–9), as found in *M. domestica*. Although the mammal-like muscle division and arrangement of the jaw adductors preceded the osteological transformation of the mandible and jaw joint[21], it was not until a later stage in mammalian evolution that further optimization to muscle function occurred. A recent study[23] using free-body analysis of the cynodont lower jaw confirmed that the musculoskeletal system was morphologically and evolutionary flexible without negatively affecting functional performance.

In our analyses, reduction in mandibular size—rather than alterations of the osteology and the muscular arrangement—produced the most notable effects on minimizing absolute jaw-joint stress. Although a decrease in size leads to two conflicting trends of reducing tensile and compressive stresses but also bite forces, stress reduction is achieved at a higher rate than bite-force reduction (exponential versus linear, respectively). Consequently, our biomechanical analyses predict that smaller mandibular size constitutes the best compromise to ameliorate loss of bite force and stress reduction in the jaw joint. This prediction is corroborated by the reduction in size in the vast majority of taxa that are phylogenetically intermediate in the cynodont–mammaliaform transition, in which such a biomechanical compromise was achieved (Figs. 1, 4b). Miniaturization has previously been discussed as a key factor during the evolution of mammals in the context of thermoregulation, nocturnality and dietary or ecological adaptations[1,2,24,25]. It has further been proposed to be a structural requirement for the acquisition of mammalian characters[1]. Although size-related stress reduction might not have been the main target of selection, it could have constituted a by-product of adaptation to a specific ecological niche that demanded small body size[26] during the early phases of radiation[4,27]. Our results demonstrate that changes to joint morphology and muscle (re-)organization have little effect on joint loading. Instead, by lowering stress and strain disproportionately to bite-force magnitude, reduction in size

© 2018 Springer Nature Limited. All rights reserved.

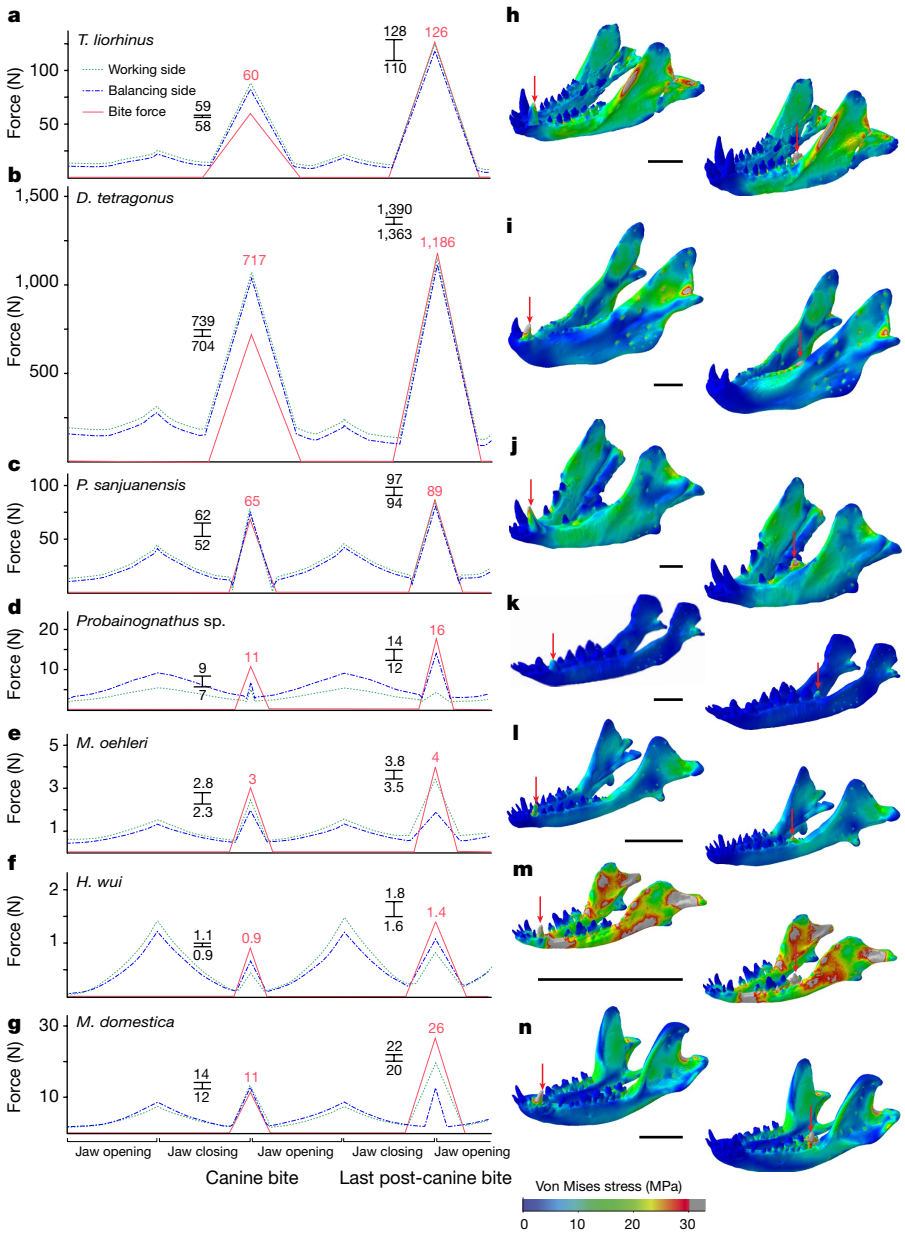


Fig. 2 | Biomechanical analysis of cynodont and mammaliaform taxa for simulated unilateral biting at canines and most-posterior tooth. **a**–**g**, Multibody dynamics analysis plots showing bite forces and joint forces (working and balancing side) during jaw opening and closing cycles. Range bars denote bite-force values obtained from the finite element models. Peak values in red represent maximum bite force obtained from multibody dynamics analysis models. **h**–**n**, Finite element von-Mises-stress contour plots for bite at canine and last tooth (indicated by red arrows). Scale bars for **h**, **j**–**n**, 10 mm; **i**, 50 mm. Sample size for each species, $n = 1$.

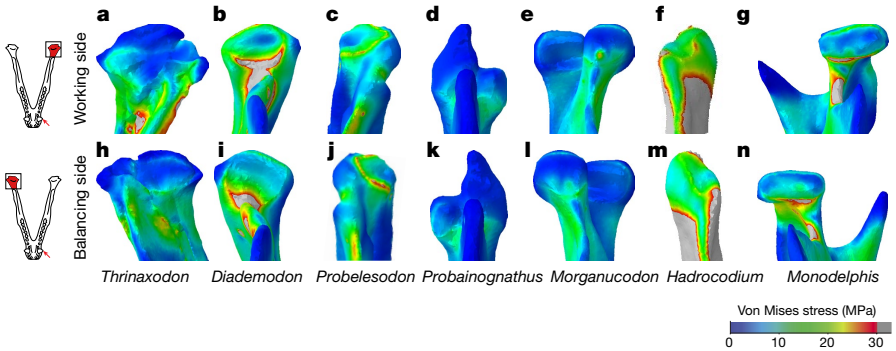


Fig. 3 | Von-Mises-stress contour plots of mandibular joint region. **a**–**g**, Jaw joint of the working side in dorsal view. **h**–**n**, Jaw joint of the balancing side in dorsal view. All contour plot images are scaled to the same size. Results are shown for a simulated unilateral bite at the most-posterior tooth.

© 2018 Springer Nature Limited. All rights reserved.

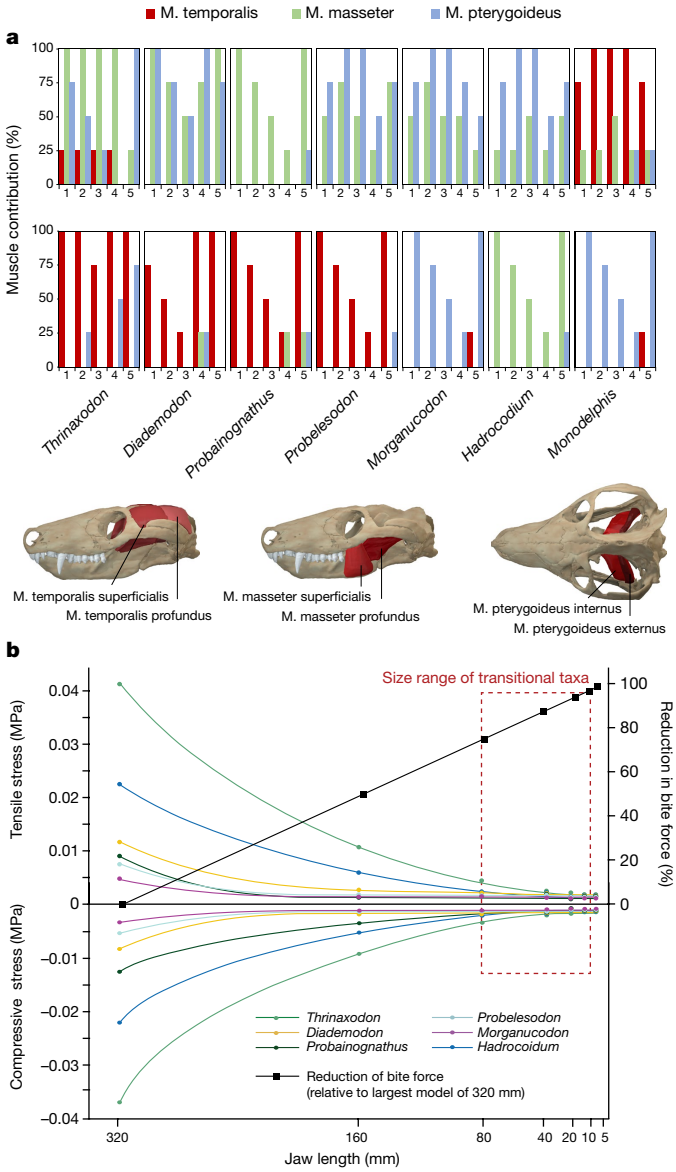


Fig. 4 | Muscle activation patterns and joint stress calculations. **a**, Muscle activation simulation to achieve the highest bite forces relative to minimum tensile (top plots) and compressive (bottom plots) stress at the jaw joint. Muscle combinations for the five highest bite force versus stress outputs shown for all taxa in decreasing order (1–5). Reconstructed adductor muscle groups are depicted in the skull images. **b**, Tensile and compressive stresses in the jaw joint and percentage reduction in bite force (relative to largest model of 320 mm) of all taxa, each scaled to seven different jaw lengths. Relative bite-force reduction is the same for all models with each successive size and is represented by a single trend line (black).

appears to be the key factor. Miniaturization of the mandibular system could therefore represent the crossing of an evolutionary Rubicon in the emergence of the temporomandibular jaw joint and in the further functional integration of postdentary bones in the middle ear, before their final separations from the mandible in derived mammaliaforms that eventually led to a spectacular diversification of crown mammals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0521-4.

Received: 15 May 2018; Accepted: 13 August 2018;
Published online 17 September 2018.

1. Kemp, T. S. *The Origin and Evolution of Mammals* (Oxford Univ. Press, Oxford, 2005).
2. Kielan-Jaworowska, Z. et al. *Mammals from the Age of Dinosaurs—Origins, Evolution, and Structure* (Columbia Univ. Press, New York, 2004).
3. Crompton, A. W. in *Studies in Vertebrate Evolution* (eds Joysey, K. A. & Kemp, T. S.) 231–253 (Oliver & Boyd, Edinburgh, 1972).
4. Luo, Z.-X. Transformation and diversification in early mammal evolution. *Nature* **450**, 1011–1019 (2007).
5. Crompton, A. W. & Hylander, W. L. in *The Ecology and Biology of Mammal-like Reptiles* (eds Hotton, N. III et al.) 263–282 (Smithsonian Institution, Washington, 1986).
6. Bramble, D. M. Origin of the mammalian feeding complex: models and mechanisms. *Paleobiology* **4**, 271–301 (1978).
7. Barghusen, H. R. in *Morphology of the Maxillomandibular Apparatus* (ed. Schumacher, G. H.) 26–32 (Georg Thieme, Leipzig, 1972).
8. DeMar, R. & Barghusen, H. R. Mechanics and the evolution of the synapsid jaw. *Evolution* **26**, 622–637 (1972).
9. Luo, Z.-X. Developmental patterns in Mesozoic evolution of mammal ears. *Annu. Rev. Ecol. Evol. Syst.* **42**, 355–380 (2011).
10. Allin, E. F. Evolution of the mammalian middle ear. *J. Morphol.* **147**, 403–437 (1975).
11. Sidor, C. A. Evolutionary trends and the origin of the mammalian lower jaw. *Paleobiology* **29**, 605–640 (2003).
12. Manley, G. A. Evolutionary paths to mammalian cochleae. *J. Assoc. Res. Otolaryngol.* **13**, 733–743 (2012).
13. Luo, Z.-X. et al. New evidence for mammaliaform ear evolution and feeding adaptation in a Jurassic ecosystem. *Nature* **548**, 326–329 (2017).
14. Reichert, C. in *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin* (ed. Müller, J.) 120–222 (W. Thome, Berlin, 1837).
15. Gaupp, E. W. T. Die Reichertsche Theorie (Hammer-, Amboss und Kieferfrage). *Archiv für Anatomie und Entwicklungsgeschichte* **1912**, 1–426 (1913).
16. Urban, D. J. et al. A new developmental mechanism for the separation of the mammalian middle ear ossicles from the jaw. *Proc. R. Soc. Lond. B* **284**, 20162416 (2017).
17. Anthwal, N., Urban, D. J., Luo, Z. X., Sears, K. E. & Tucker, A. S. Meckel's cartilage breakdown offers clues to mammalian middle ear evolution. *Nat. Ecol. Evol.* **1**, 0093 (2017).
18. Hylander, W. L. The functional significance of primate mandibular form. *J. Morphol.* **160**, 223–239 (1979).
19. Herring, S. W., Rafferty, K. L., Liu, Z. J. & Marshall, C. D. Jaw muscles and the skull in mammals: the biomechanics of mastication. *Comp. Biochem. Physiol. A* **131**, 207–219 (2001).
20. Liu, Z. J. & Herring, S. W. Bone surface strains and internal bony pressures at the jaw joint of the miniature pig during masticatory muscle contraction. *Arch. Oral Biol.* **45**, 95–112 (2000).
21. Crompton, A. W. in *Functional Morphology in Vertebrate Paleontology* (ed. Thomason, J. J.) 55–75 (Cambridge Univ. Press, Cambridge, 1995).
22. Lautenschlager, S., Gill, P., Luo, Z. X., Fagan, M. J. & Rayfield, E. J. Morphological evolution of the mammalian jaw adductor complex. *Biol. Rev. Camb. Philos. Soc.* **92**, 1910–1940 (2017).
23. Reed, D. A., Iriarte-Diaz, J. & Diekwisch, T. G. A three dimensional free body analysis describing variation in the musculoskeletal configuration of the cynodont lower jaw. *Evol. Dev.* **18**, 41–53 (2016).
24. Rowe, T. in *Mammal Phylogeny* (eds Szalay, F. S. et al.) 129–145 (Springer, New York, 1993).
25. Kemp, T. S. The origin of higher taxa: macroevolutionary processes, and the case of the mammals. *Acta Zool.* **88**, 3–22 (2007).
26. Hanken, J. & Wake, D. B. Miniaturization of body size: organismal consequences and evolutionary significance. *Annu. Rev. Ecol. Syst.* **24**, 501–519 (1993).
27. Gill, P. G. et al. Dietary specializations and diversity in feeding ecology of the earliest stem mammals. *Nature* **512**, 303–305 (2014).
28. Close, R. A., Friedman, M., Lloyd, G. T. & Benson, R. B. Evidence for a mid-Jurassic adaptive radiation in mammals. *Curr. Biol.* **25**, 2137–2142 (2015).
29. Pacheco, C. P., Martinelli, A. G., Pavanatto, A. E., Soares, M. B. & Dias-da-Silva, S. *Prozostrodon brasiliensis*, a probainognathian cynodont from the Late Triassic of Brazil: second record and improvements on its dental anatomy. *Hist. Biol.* **30**, 475–485 (2017).

Acknowledgements We thank P. Brewer, S. Chapman (Natural History Museum, London), O. Rauhut, G. Roessner (Bayerische Staatssammlung für Historische Geologie und Palaeontologie, Munich), K. Angielczyk, W. Simpson (Field Museum of Natural History, Chicago), G. Hantke and A. Kitchener (National Museums of Scotland, Edinburgh) for access to specimens in their care. T. Rowe and J. Maisano (University of Texas, Austin) generously provided digital datasets of specimens. A. Neander (University of Chicago), G. Roessner, D. Sykes (Natural History Museum London), K. Robson Brown (University of Bristol), O. Katsamenis and M. Mavrogordato (University of Southampton) assisted with computed tomography scanning. E. Ghirardello prepared the specimens and performed property testing on hedgehog mandible material. We thank J. Hopson (University of Chicago) for discussion. This work was funded by NERC (Natural Environment Research Council) grants NE/K01496X/1 (to E.J.R.) and NE/K013831/1 (to M.J.F.), and support from the University of Chicago (to Z.-X.L.).

Reviewer information *Nature* thanks C. A. Sidor and the other anonymous reviewer(s) for their contribution to the peer review of this work.

© 2018 Springer Nature Limited. All rights reserved.

Author contributions S.L., P.G.G., Z.-X.L., M.J.F. and E.J.R. conceived and designed the study. S.L., P.G., Z.-X.L. and E.J.R. arranged logistics of specimens for computed tomography scanning and collected computed tomography data. Z.-X.L. provided access to additional specimens and data. S.L. processed computed tomography data, performed digital restorations and reconstructions, and performed computational analyses. M.J.F. and E.J.R. contributed to finite element and multibody dynamics analyses. S.L., P.G., Z.-X.L., M.J.F. and E.J.R equally contributed to the analysis of results. S.L. prepared main text, figures and supplementary data. S.L., P.G., Z.-X.L., M.J.F. and E.J.R. equally contributed to editing, commenting and revising the manuscript and figures. E.J.R. and M.J.F. acquired funding.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0521-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0521-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.L. or E.J.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Specimens and digital models. Three-dimensional digital models of key cynodont and mammaliaform taxa were created for this study using the following specimens: *T. liorhinus* (NHMUK PV R 511, 511a, Natural History Museum, London), *D. tetragonus* (BSP 1934 VIII 17/2, Bayerische Staatssammlung für Historische Geologie und Paläontologie, Munich), *P. sanjuanensis* (PVSJ 411, Museo de Ciencias Naturales, Universidad Nacional de San Juan), *Probainognathus* sp. (PVSJ 410), *M. oehleri* (FMNH CUP 2320, Field Museum of Natural History, Chicago; IVPP 8685, Institute for Vertebrate Palaeontology and Palaeoanthropology), *Morganucodon watsoni* (NHMUK PV M 26144, articulated squamosal and petrosal; NHMUK PV M 92838 & M 92843, isolated quadrates; NHMUK PV M 27410, isolated fragmentary jugal), *H. wui* (IVPP 8275), *M. domestica* (National Museum of Scotland, Edinburgh). All specimens were digitized using computed tomography scanning or (as in the case of *D. tetragonus*) a photogrammetry approach. Scan methodology and settings are listed in detail in previously published methods[22]. For the model creation and the removal of taphonomic artefacts, scan data were imported into Avizo (version 8, VSG, Visualization Science Group). Datasets were segmented manually in Avizo segmentation editor to separate bone from the surrounding matrix. As all fossil specimens exhibited various preservational and taphonomic artefacts, different restoration steps were applied as previously outlined[22,30]. Detailed accounts of the restorative steps of the individual specimens are provided in the Supplementary Information and in a previous publication[22].

Three-dimensional models of the jaw adductor muscle anatomy of all fossil specimens were reconstructed digitally, following a previously described protocol[31]. Reconstructions were performed on the basis of osteological correlates that indicate muscle attachment sites. Where exact locations and boundaries between adjacent attachments were unclear, topological criteria were applied. Corresponding insertions and origins of each muscle were connected by simple point-to-point connections to evaluate the muscle arrangement and to identify possible intersections or other conflicts. Following this initial reconstruction, muscle dimensions and volumes were modelled according to spatial constraints within the bony structure. Data obtained from contrast-enhanced computed tomography scanning of *M. domestica* were consulted to further inform the fossil muscle reconstructions. Competing hypotheses regarding the exact placement and arrangement of specific muscles were evaluated by analysing muscle strain[22,32]. Full details and discussion of the reconstructed jaw adductor complex across the studied taxa have previously been published[22]. The final muscle reconstructions were used to supply input parameters for the subsequent finite element and multibody dynamics analyses. Muscle forces were calculated based on physiological cross-section area[33], which was estimated by dividing the volume of each muscle by its total length (Supplementary Table 2).

Multibody dynamics analysis. For multibody dynamics analysis, the digitally restored models of all taxa were imported into Adams (version 2013.2, MSC Software) as rigid bodies in .x_t parasolid format. The skull and jaw models were aligned manually to articulate at the quadrate–articular joint or the squamosal–dentary joint, respectively. Throughout all simulations, the skull models were kept immobile; the jaw models were allowed mobility in all degrees of freedom. Skull and jaw models were connected by spherical joint elements in Adams. Mass and inertial properties were calculated in Adams based on rigid body geometry and an average bone density of 1,764 kg/m^3[3,34]. The different adductor muscle groups were modelled as a series of spring elements that link corresponding muscle insertion and origin sites. Muscle forces were assigned according to the calculations taken from the three-dimensional reconstructions. Muscle activation was modelled by applying a dynamic geometric optimization method[35]. Unilateral and bilateral biting at the canines and the posterior-most tooth position were simulated using a rigid body box element from the Adams solids library. The box was placed perpendicular to the teeth at the aforementioned tooth positions and moved posteriorly during jaw opening phases. Bite forces and joint reaction forces for the working and balancing side joints (for the unilateral bite scenarios) were recorded throughout the bite simulations.

Two sets of simulations were performed for each taxon: (i) all models scaled to the actual size of the physical specimens and (ii) all taxa scaled to the same surface area, to evaluate the biomechanical effects of morphological differences independent of size[36]. For the latter scenario, the model of *T. liorhinus* was selected as the reference as it represents approximately the average size of all models (which range in jaw length between 13 mm and 270 mm); all other models were scaled to the same surface area as the *Thrinaxodon* model.

Finite element analysis. For finite element analysis, jaw models of all taxa were imported into Hypermesh (version 11, Altair Engineering) for the creation of solid mesh finite element models and the setting of boundary conditions. All jaw models consisted of approximately 2,500,000 tetrahedral elements. Material properties for mandibular bone and teeth were assigned based on nano-indentation results for hedgehog mandibles (bone: $E = 12$ GPa, $\upsilon = 0.30$, tooth: $E = 25.0$ GPa, $\upsilon = 0.3$); material properties for mammalian mandibular sutures ($E = 46.0$ MPa, $\upsilon = 0.35$) were taken from literature data[37]. Owing to the resolution of some computed tomography datasets, cortical and cancellous bone were not differentiated, which permitted the use of models derived from different digitization methods (computed tomography for volumetric models and photogrammetry for surface-based models). All materials were treated as isotropic and homogenous. To avoid artificially high stress and strain peaks on the articular and dentary, constraints were not directly applied to the joint region. Instead, an additional component with the same material properties was created to articulate with the joint surface. The morphology of these linking components was based on the cranial articulating joint morphology. The linking components were constrained (15 nodes on each side) from translation in the *x*, *y* and *z* directions. To simulate biting at different analogous positions, additional constraints (one node each, in the *x* and *y* directions; the *z* direction was unrestrained to allow penetration of tooth into prey) were applied to the canine and the posterior-most tooth, each for a unilateral and a bilateral scenario. Muscle forces were assigned according to the calculations taken from the three-dimensional reconstructions (Supplementary Table 2). As for the multibody dynamics analysis, a second set of simulations was performed with all models scaled to the same surface area and muscle forces scaled proportionally to analyse the models at the same relative size[36]. The models were subsequently imported into Abaqus 6.10 (Simulia) for analysis and post-processing. Biomechanical performance of the finite element models was assessed using contour plot outputs. In addition, reaction forces (equal to bite forces) at the bite points and average stress, strain and displacement values per element were obtained from the models.

For the simulation of different muscle activation patterns, load forces for the temporalis, the masseter and the pterygoideus groups were varied: each muscle group was set up to successively produce 0, 25, 50, 75 and 100% of the maximum force and all possible permutations were simulated (resulting in 125 (5^3) possible combinations, for the five different states and three muscle groups). To automate this process, an R script was used to modify the finite element analysis input files accordingly[38]. All other settings were kept constant as outlined above and analysed using Abaqus. To compare performances, bite force values and average joint stresses (von Mises, tensile and compressive) were obtained from Abaqus. Average joint stresses were calculated from 30 nodes selected in a grid pattern on the surface of the jaw joint to obtain a maximum spread and analogous point across all taxa. Results of the different muscle activation simulations were plotted in a three-dimensional coordinate system using the freely available visualization package Blender (www.blender.org) (Supplementary Figs. 7–12).

Additional finite element analysis simulations were performed for all taxa scaled to different, discrete mandible lengths: 5, 10, 20, 40, 80, 160 and 320 mm. Mandible lengths were chosen to represent the range of sizes observed across the cynodont–mammaliaform transition. Load forces were scaled for each size stage following the 3/4 power law for each taxon.

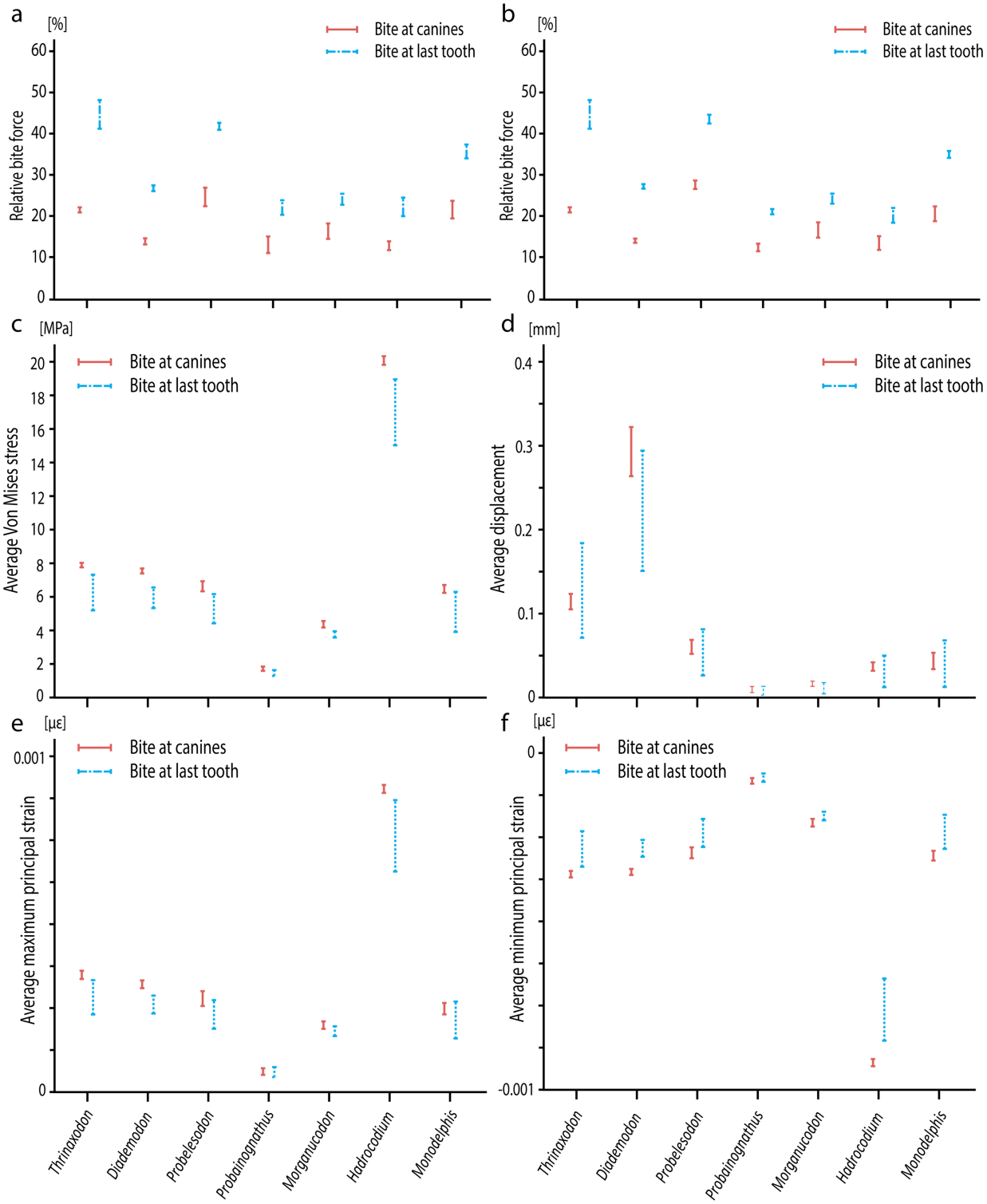
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability.

All relevant data (three-dimensional osteological, finite element analysis and multibody dynamics analysis models and computer code) are available via the DataBris repository of the University of Bristol (https://doi.org/10.5523/bris.n5f4ogftag0r2f-bffh8u7waok).

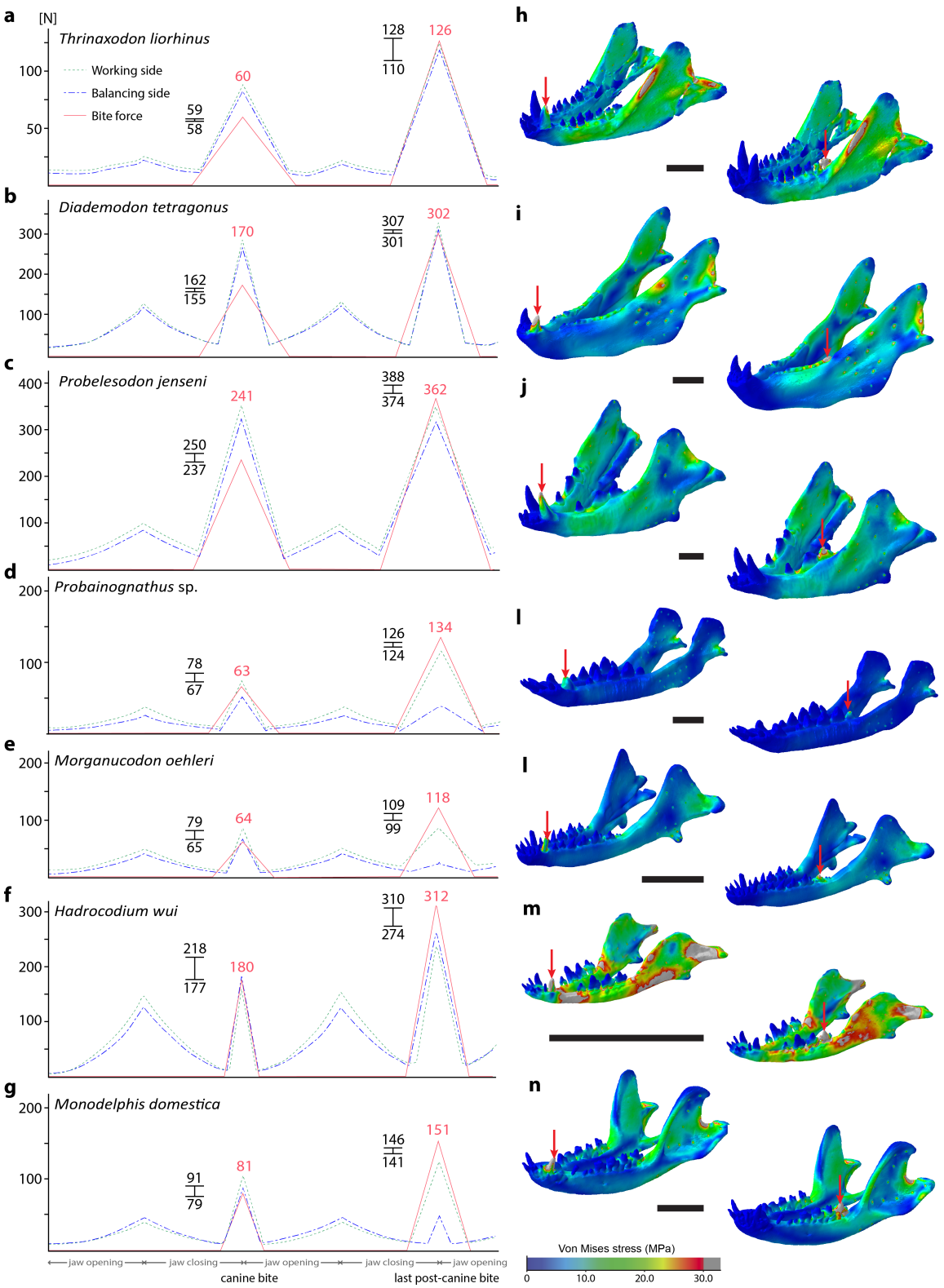
30. Lautenschlager, S. Reconstructing the past: methods and techniques for the digital restoration of fossils. *R. Soc. Open Sci.* **3**, 160342 (2016).
31. Lautenschlager, S. Cranial myology and bite force performance of *Erlikosaurus andrewsi*: a novel approach for digital muscle reconstructions. *J. Anat.* **222**, 260–272 (2013).
32. Lautenschlager, S. Estimating cranial musculoskeletal constraints in theropod dinosaurs. *R. Soc. Open Sci.* **2**, 150495 (2015).
33. Thomason, J. J. Cranial strength in relation to estimated biting forces in some mammals. *Can. J. Zool.* **69**, 2326–2333 (1991).
34. Ashman, R. B. & Rho, J. Y. Elastic modulus of trabecular bone material. *J. Biomech.* **21**, 177–181 (1988).
35. Curtis, N. et al. Predicting muscle activation patterns from motion and anatomy: modelling the skull of *Sphenodon* (Diapsida: Rhynchocephalia). *J. R. Soc. Interface* **7**, 153–160 (2010).
36. Dumont, E. R., Grosse, I. R. & Slater, G. J. Requirements for comparing the performance of finite element models of biological structures. *J. Theor. Biol.* **256**, 96–103 (2009).
37. Bright, J. A. The importance of craniofacial sutures in biomechanical finite element models of the domestic pig. *PLoS ONE* **7**, e31769 (2012).
38. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, 2017).

© 2018 Springer Nature Limited. All rights reserved.



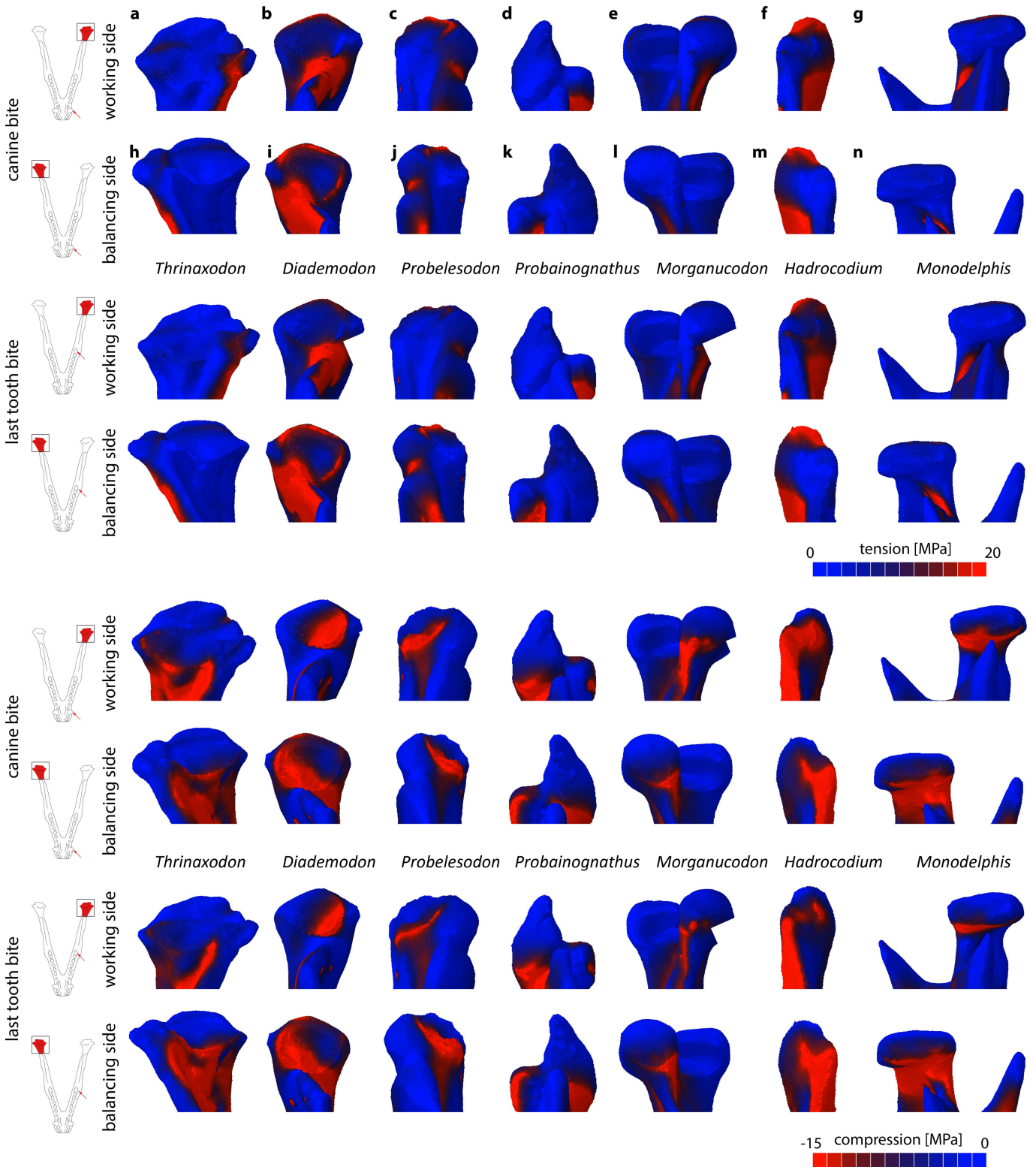
Extended Data Fig. 1 | Relative bite forces and biomechanical performance measures of cynodont and mammaliaform taxa. a, Relative bite forces for original-sized models. **b**, Relative bite forces for models scaled to the same size (with *T. liorhinus* as reference). Relative bite forces calculated as the ratio between muscle forces and resultant bite forces (obtained from reaction forces of finite element models). Range of values represents results for unilateral and bilateral bite simulations. **c**–**f**, Average per element values for Von Mises stress (**c**), displacement (**d**), maximum principal strain (**e**) and minimum principal strain (**f**). Range of values represents results for unilateral and bilateral bite simulations (for original-sized models). Sample size for each species, $n = 1$.

© 2018 Springer Nature Limited. All rights reserved.



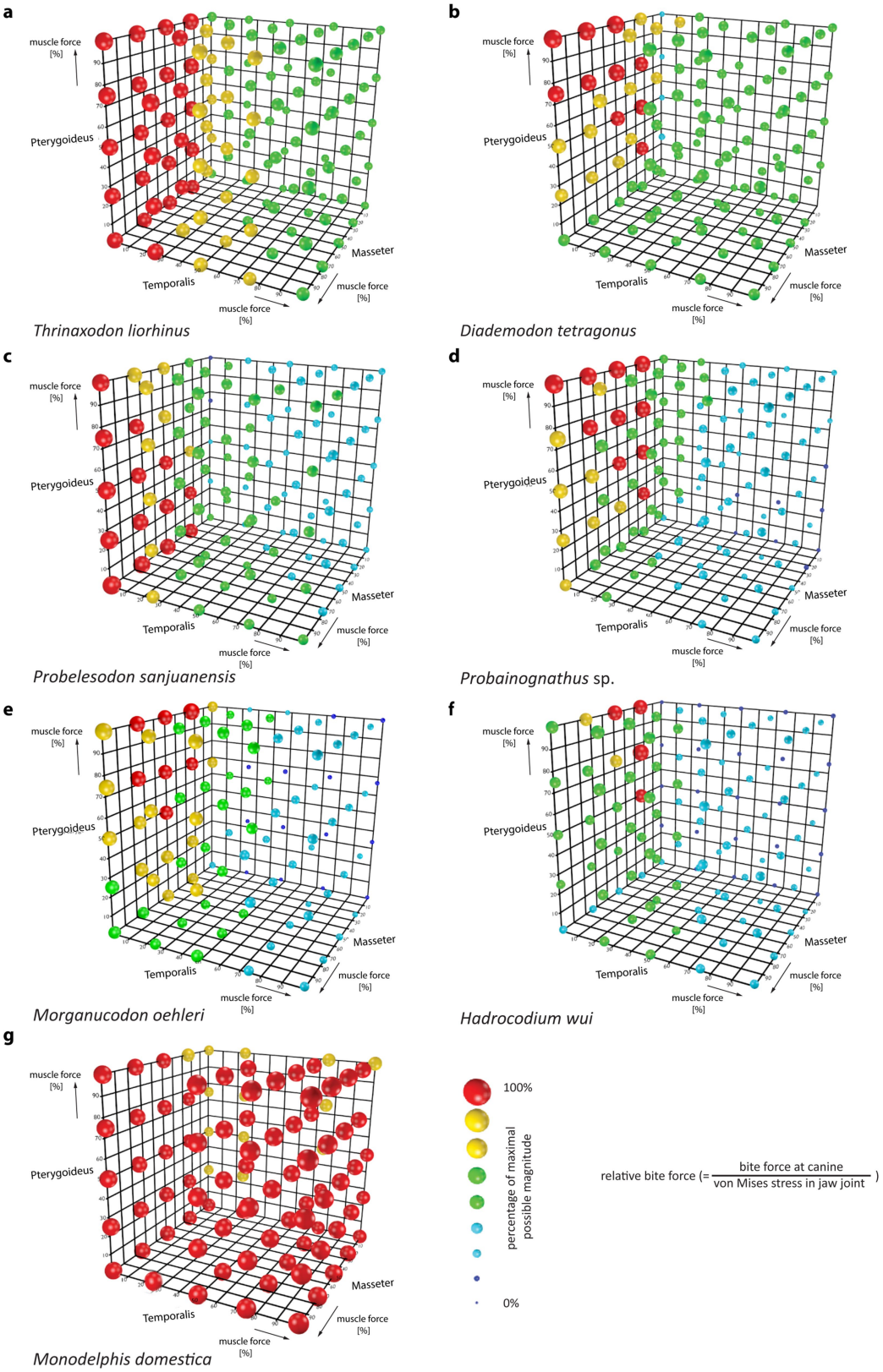
Extended Data Fig. 2 | Biomechanical analysis results of cynodont and mammaliaform taxa for simulated unilateral biting at canines and last tooth. Results for models scaled to the same size (with *T. liorhinus* as reference). **a**–**g**, Multibody dynamics analysis plots showing bite forces and joint forces (working and balancing side) during jaw opening and closing cycles. Range bars denote values obtained from reaction forces of finite element models. Peak values represent maximum bite force obtained from multibody dynamics analysis models. **h**–**n**, Finite element von-Mises-stress contour plots for bite at canine and last tooth (indicated by red arrows). Scale bars in **h**, **j**–**n**, 10 mm; **i**, 50 mm. Sample size for each species, $n = 1$.

© 2018 Springer Nature Limited. All rights reserved.



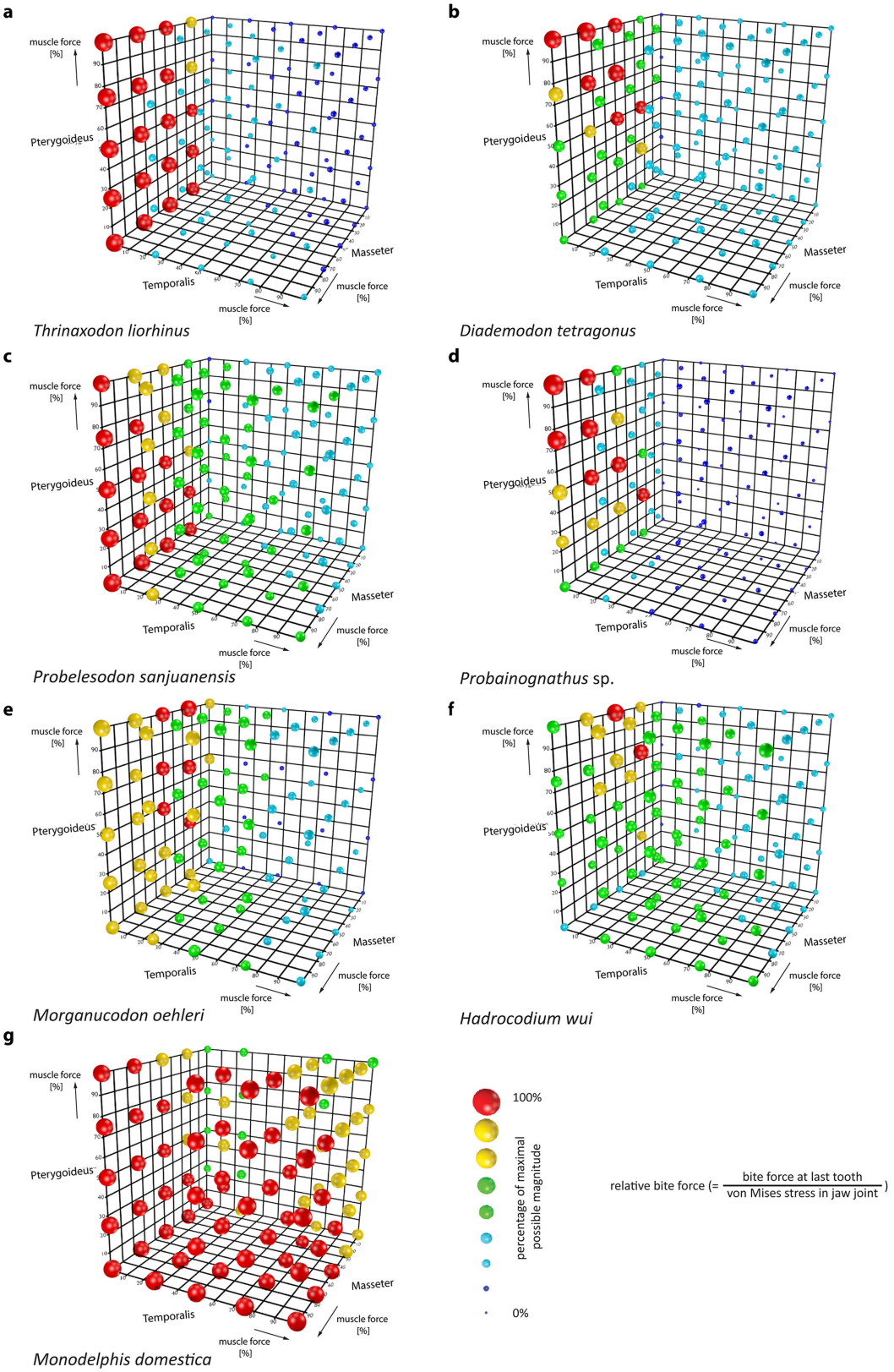
Extended Data Fig. 3 | Tensile and compressive stress contour plots of mandibular joint region. Results shown for unilateral bite at the canine (upper rows) and the last tooth position (lower rows), each for the jaw joint of the working side and the balancing side in dorsal view. All contour plot images are scaled to the same size.

© 2018 Springer Nature Limited. All rights reserved.



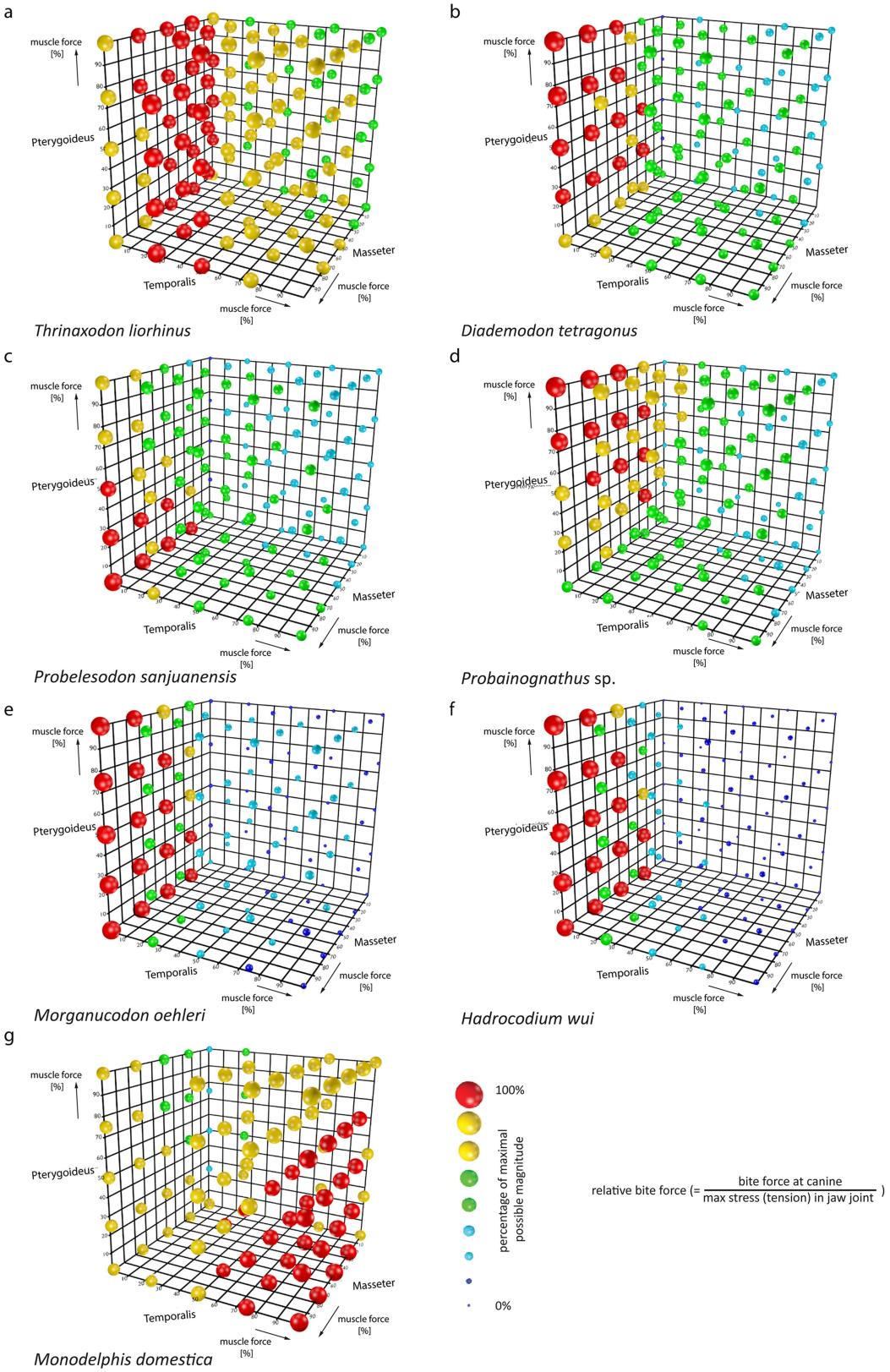
Extended Data Fig. 4 | Bite-force magnitude versus von Mises stress for different muscle activation patterns. Results shown for unilateral bite at the canine tooth position. Relative bite force measured as bite force in relation to von Mises stress occurring in the jaw joint.

© 2018 Springer Nature Limited. All rights reserved.



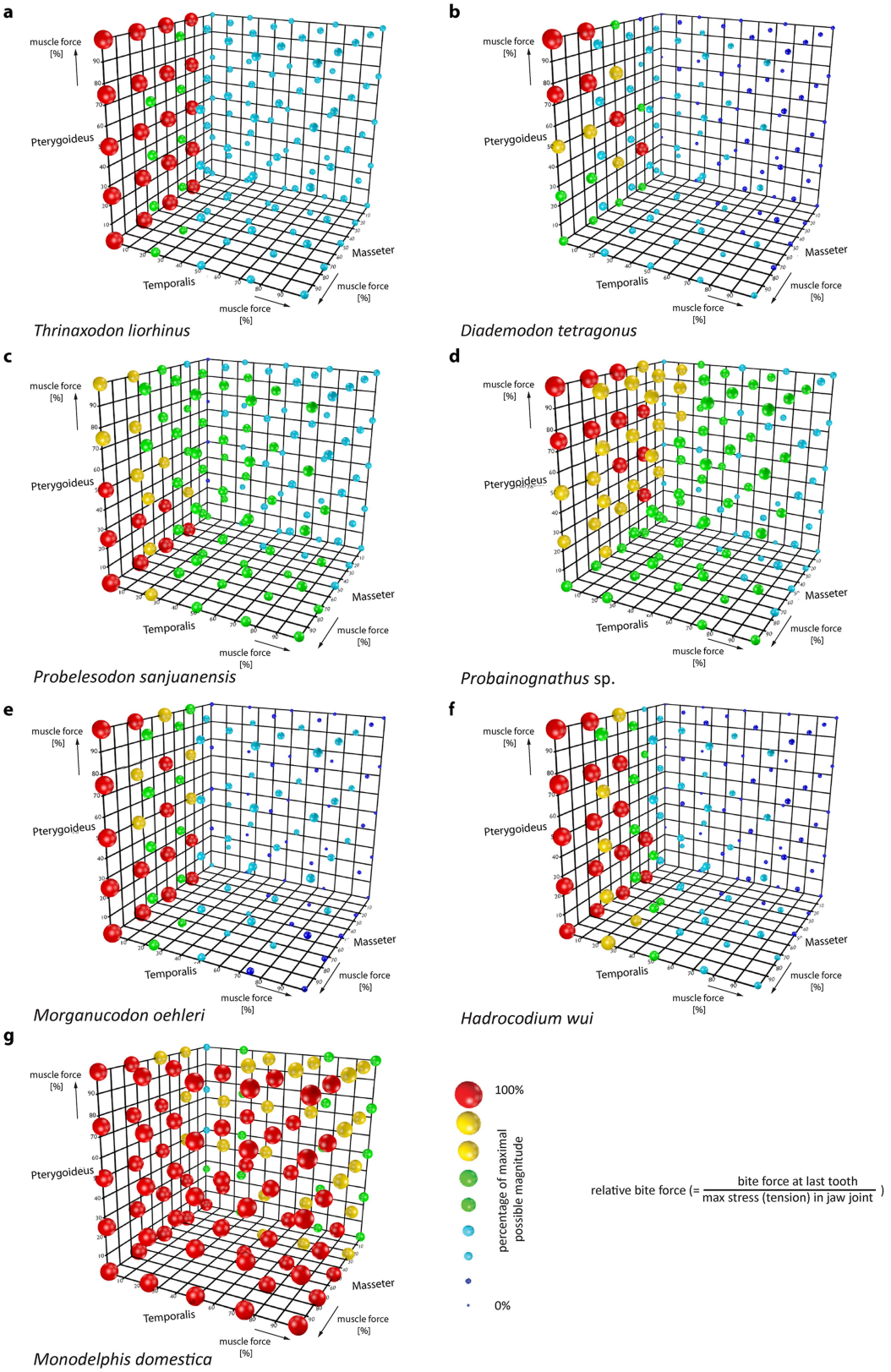
Extended Data Fig. 5 | Bite-force magnitude versus von Mises stress for different muscle activation patterns. Results shown for unilateral bite at the last tooth position. Relative bite force measured as bite force in relation to von Mises stress occurring in the jaw joint.

© 2018 Springer Nature Limited. All rights reserved.



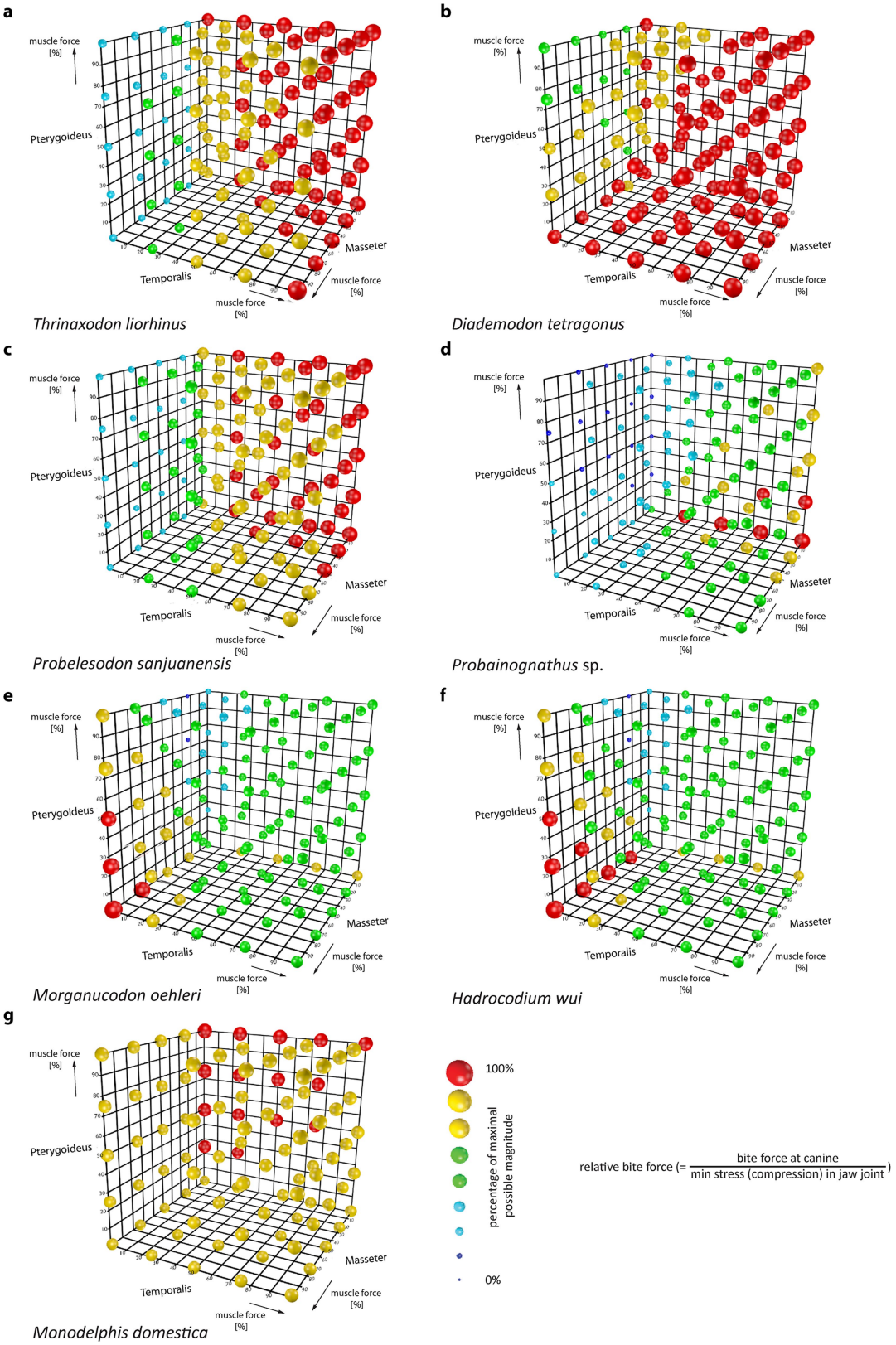
Extended Data Fig. 6 | Bite-force magnitude versus tensile stress for different muscle activation patterns. Results shown for unilateral bite at the canine tooth position. Relative bite force measured as bite force in relation to tensile stress occurring in the jaw joint.

© 2018 Springer Nature Limited. All rights reserved.



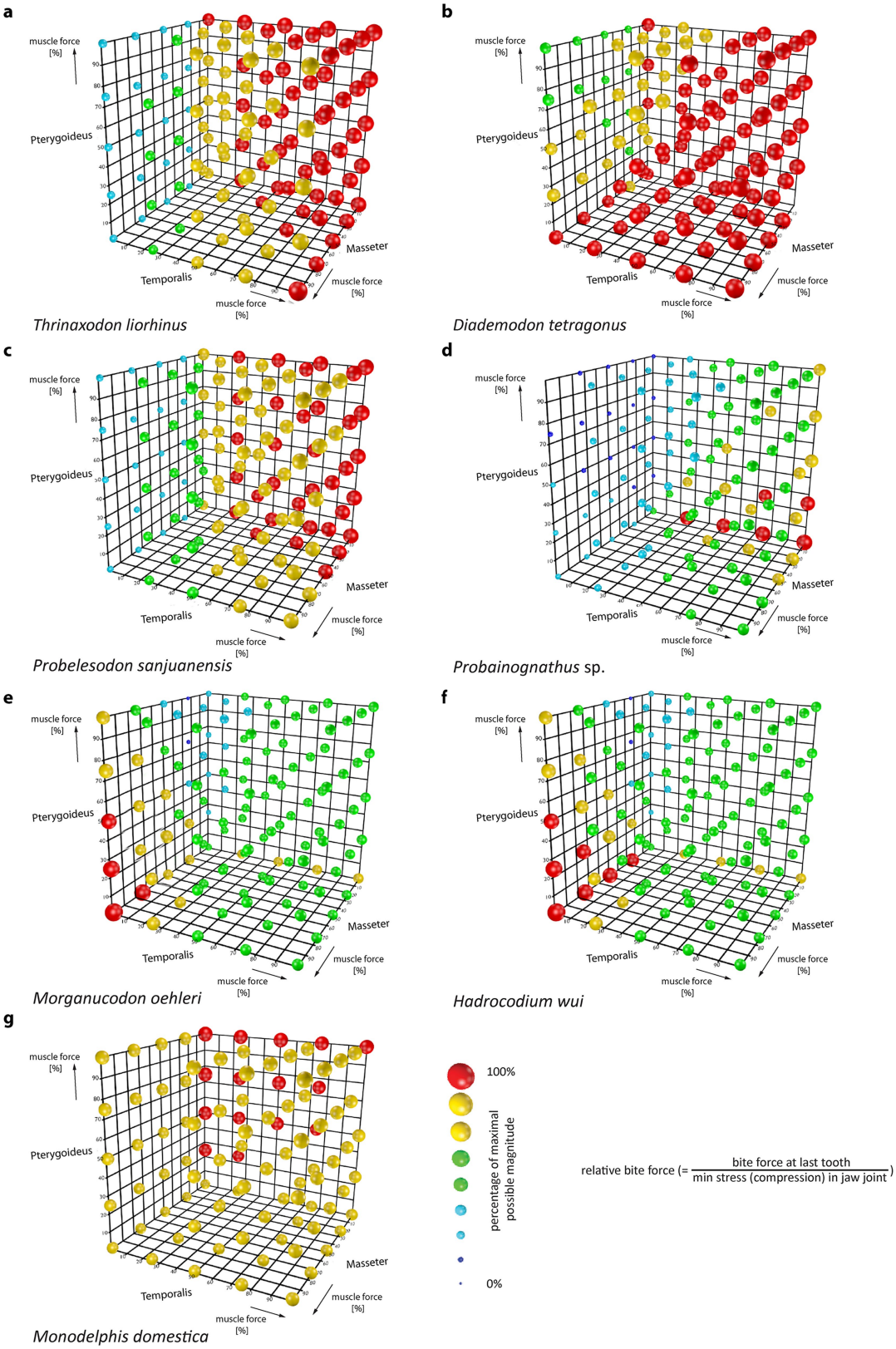
Extended Data Fig. 7 | Bite-force magnitude versus tensile stress for different muscle activation patterns. Results shown for unilateral bite at the last tooth position. Relative bite force measured as bite force in relation to tensile stress occurring in the jaw joint.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Bite-force magnitude versus compressive stress for different muscle activation patterns. Results shown for unilateral bite at the canine tooth. Relative bite force measured as bite force in relation to compressive stress occurring in the jaw joint.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Bite-force magnitude versus compressive stress for different muscle activation patterns. Results shown for unilateral bite at the last tooth. Relative bite force measured as bite force in relation to compressive stress occurring in the jaw joint.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0539-7

Hydraulic diversity of forests regulates ecosystem resilience during drought

William R. L. Anderegg[1]*, Alexandra G. Konings[2], Anna T. Trugman[1], Kailiang Yu[1], David R. Bowling[1], Robert Gabbitas[1], Daniel S. Karp[3], Stephen Pacala[4], John S. Sperry[1], Benjamin N. Sulman[5,6] & Nicole Zenes[1]

Plants influence the atmosphere through fluxes of carbon, water and energy[1], and can intensify drought through land–atmosphere feedback effects[2–4]. The diversity of plant functional traits in forests, especially physiological traits related to water (hydraulic) transport, may have a critical role in land–atmosphere feedback, particularly during drought. Here we combine 352 site-years of eddy covariance measurements from 40 forest sites, remote-sensing observations of plant water content and plant functional-trait data to test whether the diversity in plant traits affects the response of the ecosystem to drought. We find evidence that higher hydraulic diversity buffers variation in ecosystem flux during dry periods across temperate and boreal forests. Hydraulic traits were the predominant significant predictors of cross-site patterns in drought response. By contrast, standard leaf and wood traits, such as specific leaf area and wood density, had little explanatory power. Our results demonstrate that diversity in the hydraulic traits of trees mediates ecosystem resilience to drought and is likely to have an important role in future ecosystem–atmosphere feedback effects in a changing climate.

Water, carbon and energy exchanges from the land surface strongly influence the atmosphere and climate; these exchanges are dominated by plants in most ecosystems[1]. Plant physiological responses to water stress influence these fluxes[5,6], and the resulting land-surface feedback effects influence local weather as well as the regional atmospheric circulation[7]. Furthermore, changes in vegetation physiology and cover can drive shifts in sensible and latent heat fluxes that intensify droughts[2–4,8]. Anthropogenic climate change is expected to intensify the hydrological cycle globally, leading to more frequent and more severe droughts in many regions[9]. Therefore, understanding the drivers of land–atmosphere feedback effects during drought and simulating them in Earth system models is critical for robust future projections and assessment of climate change impacts.

Seminal work has shown that grassland plots with more species exhibit smaller declines in productivity during drought and recover productivity much faster following drought[10], indicating that plant biodiversity—particularly functional diversity—may be important for capturing how the land surface interacts with the atmosphere during extreme events. Indeed, it is well-established that—just as a diversified stock portfolio is more likely to survive market turbulence[11]—diversity can stabilize community function through multiple mechanisms[12]. First, diverse communities are more likely to contain species with different traits that dictate how they respond to disturbances[13]. As a result, at least some species are likely to persist through any given disturbance[14]. Second, diverse communities are more likely to contain competitors that exhibit compensatory dynamics: when drought causes one species to decline in function, its competitor may increase in function and stabilize community function[12]. Critically, diversity–stability effects are mostly absent in most global land-surface models, most of which represent each biome or plant functional type with a single set of functional traits[15], partly owing to a lack of understanding of which functional traits are the most important at ecosystem scales.

Diversity in the water transport strategies of plants has been hypothesized to play a critical part in regulating the response of an ecosystem to drought[16,17]. Plant water transport through the hydraulic continuum is mediated by a variety of traits[18,19], including xylem vulnerability to embolism, stomatal regulation and other characteristics[20]. Thus, diversity of plant hydraulic strategies and traits could buffer an ecosystem against drought, as some species will curtail gas exchange and latent heat flux through transpiration before others. However, the physiological mechanisms that govern water transport through the hydraulic continuum do not correspond well with traditional plant functional types in land-surface models[21,22]. This raises the prospect that land-surface models may be missing a critical component of functional diversity when simulating future climate scenarios with a more intense hydrological cycle.

Here we test whether plant trait diversity in forests can directly affect land–atmosphere interactions by mediating and buffering the response of latent heat fluxes to changes in water availability. We combine climate data, 352 site-years of eddy covariance data from 40 temperate and boreal forest sites across the globe (Extended Data Fig. 1 and Supplementary Table 1), and datasets of multiple plant functional traits at the species level. We test which site-level factors (for example, stand age) and plant traits are most associated with ecosystem flux variation in response to drought and whether diversity in response traits stabilizes fluctuations in ecosystem fluxes during drought periods.

We first examine which plant traits and site-level factors can best predict how eddy covariance measurements of daily latent energy exchange (a proxy for forest transpiration) vary in response to water availability. At each eddy covariance site, we quantified this drought response with two complementary metrics: the 'drought coupling' (model R^2) and the 'drought sensitivity' (standardized model coefficient values) of a multiple regression of latent energy as a function of vapour pressure deficit (VPD), soil moisture and their interaction (equation (1) in the Methods). All else being equal, forest sites with lower coupling and lower sensitivity will experience smaller variation (higher resilience) in latent energy fluxes explained by variations in VPD and soil moisture.

We considered site factors—stand age, species richness and gymnosperm fraction—and the mean and standard deviation of key functional traits expected to influence drought responses. We examined wood density, specific leaf area, maximum light-saturated photosynthetic rate (A_{max}), the water potential at which 50% of stem xylem conductivity is lost (P_{50}), the minimum stem water potential typically experienced (Psi_{min}) and several estimates of the hydraulic safety margin[18,23] (HSM; the difference between Psi_{min} and P_{50} in Fig. 1; alternative analyses in Supplementary Table 2). We compiled trait data for the dominant tree species at each site that make up more than 80% of the biomass (mean trait coverage: 91% of biomass). A strong influence of mean trait values indicates the importance of particular trait values in dictating community-wide responses to drought. By contrast, the influence of trait standard deviations suggests that it is the trait diversity per se that buffers communities[24]. We used univariate regression with

[1]School of Biological Sciences, University of Utah, Salt Lake City, UT, USA. [2]Department of Earth System Science, Stanford University, Stanford, CA, USA. [3]Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, Davis, CA, USA. [4]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [5]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA. [6]School of Natural Sciences, University of California, Merced, Merced, CA, USA. *e-mail: anderegg@utah.edu

© 2018 Springer Nature Limited. All rights reserved.

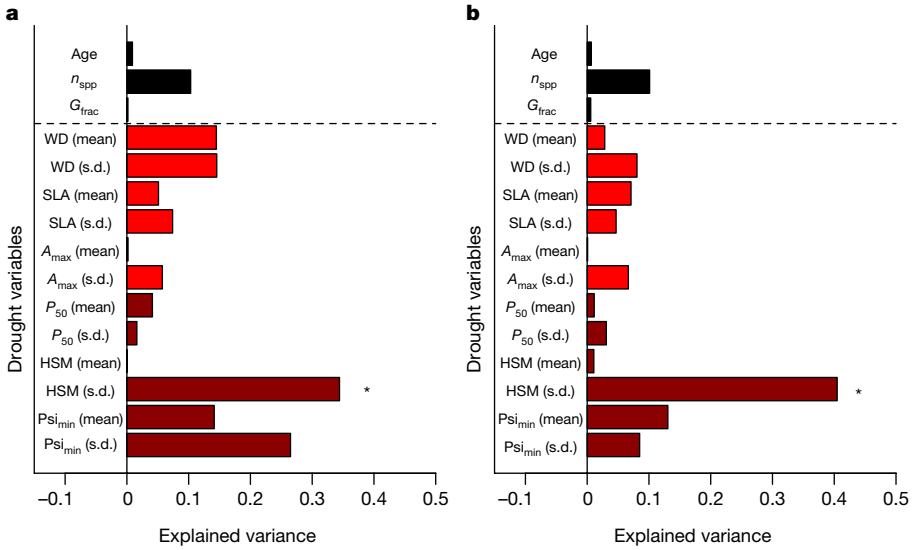


Fig. 1 | Variation in hydraulic traits mediates ecosystem flux response to drought. The percentage of explained variance (R^2) in an ordinary least-squares linear regression of community-weighted plant traits (dark red, hydraulic traits; red, other traits) and site metrics (black, above the dashed line) in explaining cross-site patterns in the coupling (**a**) and sensitivity (**b**) of latent energy exchange variation in response to drought variables. Site metrics include: forest age (Age), species richness (n_{spp}) and fraction of forest composition that is gymnosperm species (G_{frac}). Traits include the community-weighted mean and standard deviation (s.d.) of: wood density (WD), specific leaf area (SLA), A_{max}, P_{50}, HSM and Psi_{min}. For samples sizes for each trait, see Supplementary Table 3. Asterisks indicate statistically significant regressions ($P = 0.001$ (**a**) and $P = 0.02$ (**b**) after correction for multiple hypothesis testing).

individual traits, multivariate regression with model selection using the Akaike information criterion (AIC) and a machine-learning algorithm (random forests) to identify the most critical traits or site factors.

Of the tested traits, the community-level variation in HSM was the most predictive of forest flux responses to drought (Fig. 1), explaining 41% and 34% of the cross-site patterns in univariate models of drought coupling and drought sensitivity, respectively ($P < 0.01$ and $P = 0.02$ after correction for multiple hypothesis testing). However, considerable variation in drought coupling and sensitivity existed at sites where the HSM variation was low. Model selection techniques on multivariate trait models showed that the standard deviation of HSM and mean maximum photosynthetic rate were the most parsimonious models for drought sensitivity ($R^2 = 0.45$, $P < 0.001$, $\Delta AIC < -2$ from all other models). The standard deviation of HSM alone was selected as the most parsimonious model for drought coupling ($R^2 = 0.41$, $P = 0.001$). Model selection further indicated that the remaining predictors had little effect on model performance. In addition, the machine-learning algorithm consistently indicated that the standard deviation of HSM was one of the most important predictor variables (Extended Data Fig. 2).

We found that higher diversity in the HSMs of the species in an ecosystem significantly buffers ecosystem latent heat flux response to drought ($P_{sens} = 0.003$, $P_{coup} = 0.001$; Fig. 2). All else being equal, forest communities with higher hydraulic diversity (defined as a higher standard deviation in HSM) experienced smaller variation in latent energy fluxes explained by VPD and soil moisture. This pattern was generally robust, especially for the drought coupling metric, to the method of data processing including: (1) adjustment of measured latent energy exchange to correct for energy balance non-closure ($P_{coup} = 0.007$); (2) using both standardized (z-score) and raw drought variables ($P_{coup} = 0.001$); (3) alternative estimates of the HSM (Supplementary Table 2 and Extended Data Fig. 3); and (4) accounting for other potentially confounding factors such as forest age (Methods).

We found here that plant trait variation, particularly hydraulic diversity, has a critical role in the response of temperate and boreal forests to drought. The importance of trait variation metrics (Fig. 1), as opposed to trait mean metrics, highlights a critical role of compensatory dynamics among species in how functional diversity mediates ecosystem responses to climate extremes (Supplementary Information). We acknowledge, however, that our analysis is limited to 40 flux sites and the examined traits do not capture all the important elements of the drought-response strategies of each species. More traits, particularly rooting depth and allometric differences (for example, leaf-to-sapwood ratios), will probably help to explain some of the remaining variations in ecosystem resilience. Furthermore, our sites generally do not include droughts that are severe enough to trigger tree mortality or changes in species composition, which may be important processes that influence the long-term response of an ecosystem to drought[25].

Recent advances have shown that plant hydraulics can be used to predict stomatal responses to changing environmental conditions[19,26] and the simulation of plant hydraulic transport in Earth-system models

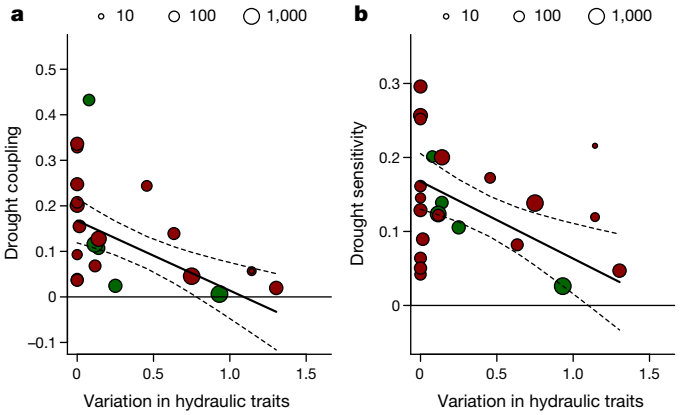


Fig. 2 | Ecosystem sensitivity to drought as a function of community variation in hydraulic safety margin. a, Drought coupling is expressed as the percentage of explained variation (R^2). **b**, Drought sensitivity is expressed as the summed absolute values of standardized coefficients of drought variables regressed against latent energy (LE) exchange as a function of daily VPD, soil moisture (SM) and their interaction (regression: $LE = f(VPD, SM, VPD \times SM)$). Hydraulic variation is expressed as the community-weighted standard deviation in the HSM of each species. Colours indicate biomes of deciduous broadleaf (green) and needleleaf (red) forests. The size of the dot indicates the number of days included for each flux site. The solid black line is the best-fit ordinary least-squares linear regression and dashed lines are the 95% confidence interval of the regression fit ($n = 23$ independent sites).

© 2018 Springer Nature Limited. All rights reserved.

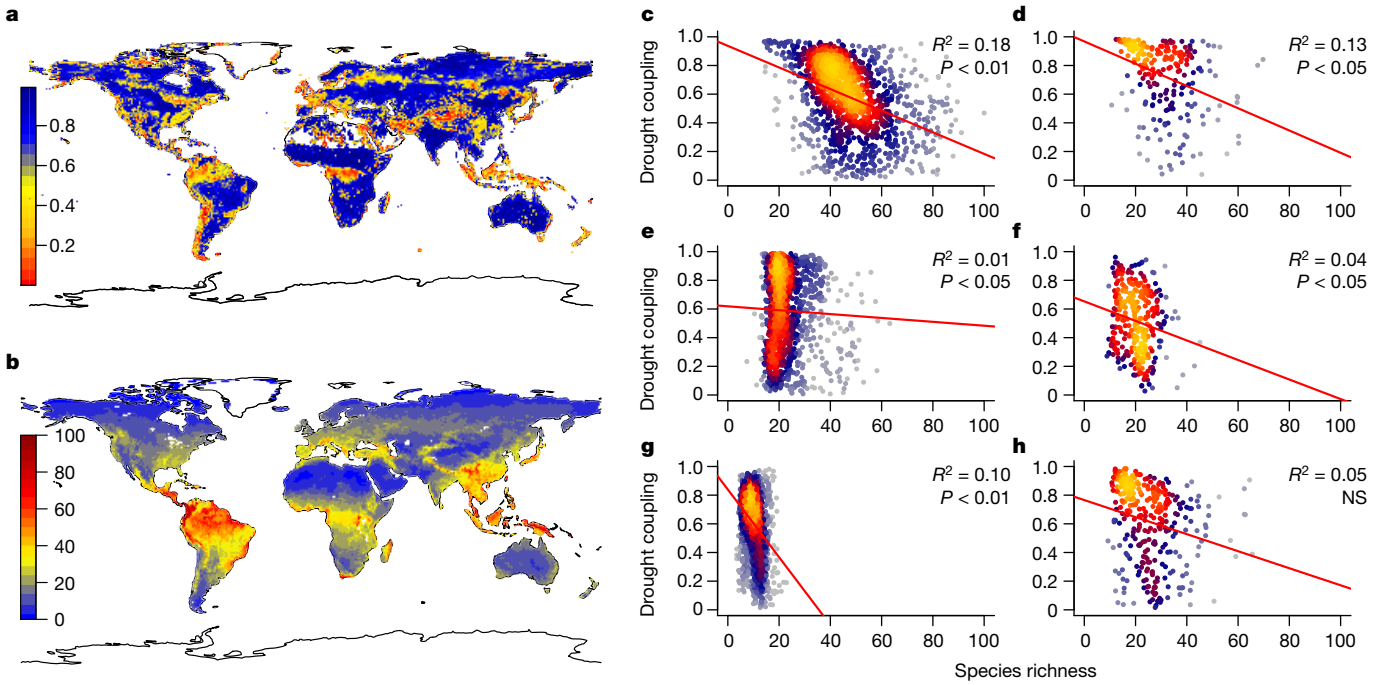


Fig. 3 | Forest ecosystem response to drought estimated from remote-sensing-derived vegetation water content variation is influenced by species richness. a, Drought coupling as the percentage of explained variation (R^2) in an ordinary least-squares linear regression by drought variables on an index of aboveground plant water content variation at midday (regression: vegetation optical depth at midday (VOD_{midday}) = f(VOD_{night}, VPD). **b**, Native plant species richness (percentage of the maximum). Data available from http://ecotope.org/anthromes/biodiversity/plants/data/. **c**–**h**, Regressions between these two variables for six major biomes. **c**, Tropical and subtropical moist broadleaf forests. n = 1,380 grid cells. **d**, Tropical and subtropical dry broadleaf forests. n = 241 grid cells. **e**, Temperate broadleaf and mixed forests. n = 1,289 grid cells. **f**, Temperate coniferous forests. n = 318 grid cells. **g**, Boreal forests. n = 1,784 grid cells. **h**, Mediterranean-type forests, woodlands and shrubs. n = 319 grid cells. Each point represents an individual grid cell from the map and redder colours indicate a higher density of points. Red lines show ordinary least-squares regression lines of best fit. Numbers in the upper right of panels indicate the linear generalized least-squares regression R^2 and P values indicate statistical significance of that regression after accounting for spatial autocorrelation.

is now possible[27]. Current land-surface models largely do not simulate diversity in strategies in plant response to drought, nor do they currently include plant hydraulics, potentially leading to substantial carbon cycle uncertainty[28]. Efforts are underway to develop and test functional trait-based models at the Earth-system scale[15]. Incorporating plant hydraulic diversity, especially through diversity in P_{50} and stomatal response to water potential (for example, HSM), will also improve the representation of related processes. For example, the loss of plant hydraulic transport is the key mechanism of drought-induced mortality[29] and can predict mortality risk among species in diverse communities[30].

Our broadly distributed temperate and boreal forest flux tower analyses suggest that drought resilience signals from biodiversity might be observable in satellite-based estimates of vegetation water content (Methods and Extended Data Figs. 4, 5). As a preliminary exploration, we examined covariation of the variability in vegetation water content with dominant tree species diversity across the continental United States and conducted a similar, more speculative analysis at global scales. Similar to the eddy flux tower findings, higher species diversity was associated with a more buffered response of vegetation water content to drought indicators across the United States (P_{sens} = 0.0005; P_{coup} < 0.0001; Extended Data Fig. 6) and globally within tropical moist (P < 0.0001), tropical dry (P = 0.03), temperate broadleaf (P = 0.01), temperate conifer (P = 0.02) and boreal (P < 0.0001) forest ecosystems, and was marginally associated in Mediterranean-type woodlands (P = 0.07) (Fig. 3 and Extended Data Fig. 7). Species richness was also an important predictor variable identified using the machine-learning algorithm (Extended Data Fig. 8), for which the complete model explained 27% of the variance in drought responses in global forests. The influence of diversity on ecosystem sensitivity showed strong biome-specific differences (Fig. 3). Across biomes, we observed a saturating relationship between the biodiversity and drought sensitivity of the ecosystems (Extended Data Fig. 9), although this appears to be driven by biome-level differences in sensitivity. Although this initial exploration should be treated with caution before a more detailed observational validation is performed, it provides a starting point for global analyses of how forest diversity regulates ecosystem resilience during drought that is consistent with conclusions drawn from flux tower analysis.

We have documented a fundamental effect of trait variation on ecosystem stability that directly influences the atmosphere and climate system. Temperate and boreal forest ecosystems with higher hydraulic diversity are more buffered to changing drought conditions. Owing to the paucity of eddy covariance sites and physiological trait measurements at tropical forest sites, additional measurements are needed to test these diversity–stability patterns in more diverse forests. Our initial analysis suggests that satellite measurements of canopy water content may be promising for overcoming some of the scarcity barrier of limited eddy covariance sites for scaling and testing drought responses at continental scales. Our results provide evidence that hydraulic diversity is a critical element of biodiversity for next-generation land-surface models to include to improve simulations of carbon, water and energy fluxes in a rapidly changing climate.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0539-7.

Received: 16 November 2017; Accepted: 14 August 2018;
Published online 19 September 2018.

1. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320**, 1444–1449 (2008).
2. Seneviratne, S. I. et al. Impact of soil moisture–climate feedbacks on CMIP5 projections: first results from the GLACE-CMIP5 experiment. *Geophys. Res. Lett.* **40**, 5212–5217 (2013).

© 2018 Springer Nature Limited. All rights reserved.

3. Berg, A. et al. Land–atmosphere feedbacks amplify aridity increase over land under global warming. *Nat. Clim. Change* **6**, 869–874 (2016).
4. Dirmeyer, P. A. Vegetation stress as a feedback mechanism in midlatitude drought. *J. Clim.* **7**, 1463–1483 (1994).
5. Sulman, B. N. et al. High atmospheric demand for water can limit forest carbon uptake and transpiration as severely as dry soil. *Geophys. Res. Lett.* **43**, 9686–9695 (2016).
6. Novick, K. A. et al. The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nat. Clim. Change* **6**, 1023–1027 (2016).
7. Khanna, J., Medvigy, D., Fueglistaler, S. & Walko, R. Regional dry-season climate changes due to three decades of Amazonian deforestation. *Nat. Clim. Change* **7**, 200–204 (2017).
8. Teuling, A. J. et al. Contrasting response of European forest and grassland energy exchange to heatwaves. *Nat. Geosci.* **3**, 722–727 (2010).
9. IPCC. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. (Cambridge Univ. Press, Cambridge, 2012).
10. Tilman, D. Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
11. Doak, D. F. et al. The statistical inevitability of stability–diversity relationships in community ecology. *Am. Nat.* **151**, 264–276 (1998).
12. Tilman, D., Isbell, F. & Cowles, J. M. Biodiversity and ecosystem functioning. *Annu. Rev. Ecol. Evol. Syst.* **45**, 471–493 (2014).
13. Naeem, S. & Wright, J. P. Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. *Ecol. Lett.* **6**, 567–579 (2003).
14. Mori, A. S., Furukawa, T. & Sasaki, T. Response diversity determines the resilience of ecosystems to environmental change. *Biol. Rev. Camb. Philos. Soc.* **88**, 349–364 (2013).
15. Scheiter, S., Langan, L. & Higgins, S. I. Next-generation dynamic global vegetation models: learning from community ecology. *New Phytol.* **198**, 957–969 (2013).
16. Matheny, A. M., Mirfenderesgi, G. & Bohrer, G. Trait-based representation of hydrological functional properties of plants in weather and ecosystem models. *Plant Divers.* **39**, 1–12 (2017).
17. Anderegg, W. R. L. et al. Woody plants optimise stomatal behaviour relative to hydraulic risk. *Ecol. Lett.* **21**, 968–977 (2018).
18. Choat, B. et al. Global convergence in the vulnerability of forests to drought. *Nature* **491**, 752–755 (2012).
19. Sperry, J. S. et al. Pragmatic hydraulic theory predicts stomatal responses to climatic water deficits. *New Phytol.* **212**, 577–589 (2016).
20. Brodribb, T. J. Xylem hydraulic physiology: the functional backbone of terrestrial plant productivity. *Plant Sci.* **177**, 245–251 (2009).
21. Anderegg, W. R. Spatial and temporal variation in plant hydraulic traits and their relevance for climate change impacts on vegetation. *New Phytol.* **205**, 1008–1014 (2015).
22. Konings, A. G. & Gentine, P. Global variations in ecosystem-scale isohydricity. *Glob. Change Biol.* **23**, 891–905 (2017).
23. Martin-StPaul, N., Delzon, S. & Cochard, H. Plant resistance to drought depends on timely stomatal closure. *Ecol. Lett.* **20**, 1437–1447 (2017).
24. Suding, K. N. et al. Scaling environmental change through the community-level: a trait-based response-and-effect framework for plants. *Glob. Change Biol.* **14**, 1125–1140 (2008).
25. Zhang, T., Niinemets, Ü., Sheffield, J. & Lichstein, J. W. Shifts in tree functional composition amplify the response of forest biomass to climate. *Nature* **556**, 99–102 (2018).
26. Wolf, A., Anderegg, W. R. L. & Pacala, S. W. Optimal stomatal behavior with competition for water and risk of hydraulic impairment. *Proc. Natl Acad. Sci. USA* **113**, E7222–E7230 (2016).
27. Bonan, G. B., Williams, M., Fisher, R. A. & Oleson, K. W. Modeling stomatal conductance in the earth system: linking leaf water-use efficiency and water transport along the soil–plant–atmosphere continuum. *Geosci. Model Dev.* **7**, 2193–2222 (2014).
28. Trugman, A., Medvigy, D., Mankin, J. & Anderegg, W. R. L. Soil moisture stress as a major driver of carbon cycle uncertainty. *Geophys. Res. Lett.* **45**, 6495–6503 (2018).
29. Nardini, A., Battistuzzo, M. & Savi, T. Shoot desiccation and hydraulic failure in temperate woody angiosperms during an extreme summer drought. *New Phytol.* **200**, 322–329 (2013).
30. Anderegg, W. R. L. et al. Meta-analysis reveals that hydraulic traits explain cross-species patterns of drought-induced tree mortality across the globe. *Proc. Natl Acad. Sci. USA* **113**, 5024–5029 (2016).

Acknowledgements We thank M. Beninati, R. O'Dell and J. Gallafent for assistance with trait compilation. W.R.L.A. acknowledges funding from the University of Utah Global Change and Sustainability Center, NSF Grant 1714972 and 1802880 and the USDA National Institute of Food and Agriculture, Agricultural and Food Research Initiative Competitive Programme, Ecosystem Services and Agro-ecosystem Management, grant no. 2018-67019-27850. A.T.T. acknowledges funding from USDA National Institute of Food and Agriculture Postdoctoral Research Fellowship grant no. 2017-07164. This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center and the OzFlux, ChinaFlux and AsiaFlux offices.

Reviewer information *Nature* thanks D. Baldocchi, S. Delzon, S. Jansen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.R.L.A. designed the study with all authors providing input. A.G.K., K.Y., R.G. and N.Z. contributed data and assisted with data collection. W.R.L.A., A.G.K. and A.T.T. analysed the data. W.R.L.A. wrote the paper with all authors providing input.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0539-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0539-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to W.R.L.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Flux tower analysis. We used the FLUXNET2015 Tier 1 dataset of eddy covariance sites (http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/) around the world to quantify ecosystem sensitivity to drought and the degree to which plant traits mediate ecosystem flux variation due to changes in environmental drivers. The FLUXNET2015 dataset provides a standardized set of fluxes of carbon, water and energy at over 210 sites and has undergone a standard set of quality assurance and quality control tests and gap-filling[31]. Using the biome classification from the International Geosphere–Biosphere Programme (IGBP) provided for the FLUXNET2015 sites, we selected sites that were forest or woodland ecosystems (deciduous broadleaf forest, evergreen needleleaf forest, evergreen broadleaf forest or mixed forest) and sites that had not experienced disturbance in at least 10 years before flux monitoring, to avoid sites with frequent disturbances or rapidly developing vegetation. This led to a list of 65 candidate sites. Because adequate on-site climate data are needed to determine drought sensitivity and knowledge of the functional traits of dominant canopy species, we further excluded sites that did not have on-site measurements of the vapour pressure saturation deficit of air (VPD, calculated from temperature and relative humidity and reported as a derived variable in the FLUXNET2015 dataset), soil water content of the top 30 cm, or sites that did not have adequate species coverage of functional trait measurements (see below) for at least two of our functional traits. This led to a final site list of 40 forest sites around the world (Supplementary Table 1 and Extended Data Fig. 1), covering 352 site-years of data.

We used a previously published method[5] that was specifically designed to estimate drought sensitivity at a given eddy covariance site by focusing on days when water availability was most likely to control ecosystem fluxes (for example, screening out days when other meteorological drivers, such as radiation, were likely to be more important influences on fluxes). In brief, we used the daily-level records of latent energy exchange as dependent variables to quantify variation in water and energy fluxes. Latent energy is dominated by plant transpiration in forest ecosystems under certain conditions[5,6] and thus variation in latent energy exchange is a major mechanism through which plants on the land surface affect the atmosphere. Following previous analyses[5], we conservatively selected only days on which latent heat was dominated by transpiration, changes in leaf area were likely to be relatively minor, and flux variation was likely to be influenced by water availability. Thus, at each site, we restricted our analysis to days during the peak of growing season, quantified as days on which: (1) the average temperature was above 15 °C; (2) solar radiation was high, quantified as days in which 24-h average photosynthetic photon flux density was greater than 500 μmol m^{-2} s^{-1}; and (3) sufficient evaporative demand existed to drive water fluxes, quantified as 24-h average VPD > 0.5 kPa. For all analyses, we used both the 'LE' and 'LE.CORR' variables reported by the FLUXNET2015 database for latent energy exchange. LE.CORR reflects a correction factor, which assumes that the measured Bowen ratio is correct; this factor scales the energy fluxes with the measured Bowen ratio and is used to force energy balance closure at each flux site. Our results were robust to either variable. Following previous analyses[5], we standardized the timeseries of latent energy by dividing by the mean latent energy at a given site where VPD was between 0.9 and 1.1 kPa and volumetric soil water content was >90th percentile (considered to be well-hydrated conditions). This standardization allowed for cross-site comparison of latent energy sensitivity to climate conditions, by accounting for fixed extrinsic (for example, stand biomass or leaf area) drivers.

Once this subset of days was determined, we constructed a multivariate drought model for each site by performing a multiple linear regression of daily latent energy exchange (LE) as a function of daily VPD, soil moisture (SM) and their interaction, as has previously been done for cross-site comparisons in eddy covariance data[5,6]:

$$\mathrm{LE} = \beta_1 \mathrm{VPD} + \beta_2 \mathrm{SM} + \beta_3 \mathrm{SM} \times \mathrm{VPD} + \varepsilon \quad (1)$$

in which β_i indicates the regression coefficient for term i. Consistent with previous drought regression studies that focused on temperate forest eddy covariance sites across the United States[5,6], soil moisture in the top 30 cm was used because data for soil moisture from deeper layers are rarely available. We note that soil moisture is challenging to quantify across flux sites and absolute values of volumetric water content of soil should be treated with caution. This multivariate regression method has been used successfully in multiple temperate forest studies across the United States to quantify the sensitivity of ecosystems to the critical components of drought stress[5,6] and performed relatively well at our flux sites (Extended Data Fig. 10).

The adjusted R^2 values of these regressions represent the degree to which ecosystem-level variation in latent energy exchange was controlled by VPD and soil moisture and was subsequently used as a dependent variable ('drought coupling' in the main text) in our cross-site analysis. In addition, we further performed all multiple linear regressions with standardized (z-score) VPD, soil moisture and their interaction, which enabled quantitative comparison of the regression coefficients[32]. We summed the absolute values of the three coefficients to provide a metric of 'drought sensitivity' at each flux site.

From published literature, we compiled the species composition and, where possible, dominance of tree species at each flux site (Supplementary Tables 3, 4). Using a combination of the metadata and studies listed on the host site for each flux tower and targeted literature searches, we created a database of the dominant tree species in the footprint of each flux tower. For 22 of the 40 sites (Supplementary Tables 3, 4), we were able to acquire estimates of relative dominance of each species, typically through metrics of basal area or composition percentage, and we converted these metrics to the proportion of each species within the total plant community (that is, between 0 and 1). These values were used in a subsequent analysis to calculate community-weighted (for example, dominance-weighted) trait means and standard deviations.

Trait analysis. We compiled trait data for dominant tree species that comprised >80% of the biomass and/or composition at all sites (mean across all sites: 91%), which enabled capturing first-order effects of functional traits on ecosystem water fluxes. For sites for which biomass and/or composition data were not available, we assumed that all 'dominant' species listed in the publication had equal composition. We used the Global Wood Density Database[33,34] to compile wood density data for species. We used a previously published dataset[35] to compile the traits of each species for light-saturated maximum photosynthetic rate and specific leaf area (SLA). Traits related to photosynthetic rates and SLA are prominent traits that are typically considered as inputs for next-generation 'trait-based' models, and SLA is often related to leaf metabolic rates and lifespan in these models. We used the Xylem Functional Traits database[36] to compile the water potential at 50% loss of hydraulic conductivity (P_{50}) and the HSM, defined as the difference between P_{50} and the minimum water potential experienced (see alternative definitions in Supplementary Table 2), for each species. The HSM has previously been observed to be a critical predictive trait[18], because it provides an integrative assessment of species 'riskiness' during drought, integrating both hydraulic vulnerability to water potential and stomatal responses[23] and/or potentially diversity in the rooting strategies of each species. We then calculated the mean and standard deviation of each trait for each site as the community-weighted mean and standard deviation. Supplementary Table 3 describes the proportion of species composition at each site for which we had adequate trait data and sites were dropped from trait-level analyses if they had insufficient trait data.

Using the drought correlation and drought sensitivity of flux sites as dependent variables, we used three complementary approaches to determine which site factors and functional traits were most important. First, we performed univariate ordinary least-squares regressions between functional traits and the drought response metrics (Fig. 1). Because this involved multiple hypothesis testing, we implemented the Bonferroni correction to adjust P values for multiple comparisons. Because the hydraulic trait P_{50} can be sensitive to method artefacts, we repeated this analysis with alternative values for the few (n = around 6–8) potentially problematic species (Supplementary Table 5) and observed that our results were robust. Second, we used models of multiple traits, determining the most parsimonious model through model selection procedures. Because variable importance estimates can be biased by collinear predictor variables, we used a matrix of pairwise correlations and removed any variable with high correlations ($R > 0.5$) with other predictor variables. Each pairwise correlation was performed and the variable with the lower correlation with the dependent variable was removed[37]. With this reduced set of predictor variables, we then used both forward and backward stepwise model selection via AIC, with ΔAIC values of < -2 used as a criterion to drop variables[38]. This technique provides a rigorous estimate of the most parsimonious model, and the coefficients of the (previously standardized to z-score) predictor variables that remain can be compared directly. Because the quantity of data available across sites was highly variable (range 10–1,200 days that met the above criterion) and thus the variable statistical insight that can be drawn from each site varied based on this sample size (that is, a regression that covers 20 days should be down-weighted relative to a regression that covers 800 days), we weighted each site by the number of days for which data were available. For plotting Fig. 2, one site with adequate HSM data was classified 'mixed forest' by the FLUXNET2015 database and for simplicity we plotted it as an 'evergreen needleaf forest', because it was dominated by gymnosperm species.

Finally, we used a machine-learning algorithm of random forests[39] to estimate the variable importance on the non-collinear traits identified as above. We chose this algorithm because it (1) performs well among machine-learning algorithms; (2) makes no assumptions about the distribution (for example, normality) of the input data; (3) makes no assumptions on the functional form of the relationship between independent and dependent variables (for example, linear, nonlinear, and so on); and (4) can handle interactions between independent variables. We examined variable importance using the total decrease in node impurities for each variable.

Satellite estimates of drought-driven variation in vegetation water content. We tested spatial patterns of variations in drought sensitivity using ecosystem-scale

© 2018 Springer Nature Limited. All rights reserved.

estimates of vegetation water content from microwave radiometry-derived vegetation optical depth (VOD). We used VOD from the land parameter data record, which was retrieved from the X-band (10.7 GHz) observations of the advanced microwave scanning radiometer–earth observing system (AMSR-E) based on simultaneous retrieval of atmospheric water vapour, surface soil moisture, VOD and the canopy scattering albedo. Measurements across a variety of ecosystem types have previously shown that VOD is proportional to plant water content[40] and therefore also to leaf water potential[41], although additional research is needed to determine how the relationship between VOD and plant water potential varies across ecosystems. Full retrieval algorithm details can be found in previously published studies[42–44]. Rainy days were filtered from the record to avoid contamination from intercepted water on the leaves affecting the AMSR-E observations. Data for which the land-surface temperature (derived from higher-frequencies of the AMSR-E radiometer) was below 273 K were assumed to represent frozen soil conditions for which the VOD retrieval algorithms are not valid, and were also removed from the dataset. We used data from January 2003–December 2010 (the AMSR-E failed in 2011). Rain was determined based on data from the Global Precipitation Climatology Project[45]. We compared daily midday VOD estimates to the total daily latent energy fluxes at our flux tower sites and found that the two were linearly related despite each VOD observation representing an area 2–3 orders of magnitude greater than the flux tower fetch ($P < 0.0001$, Extended Data Fig. 4), indicating that both captured the broad-scale ecosystem response to water availability. Furthermore, we observed that hydraulic diversity was correlated with tree species richness in the United States ($R^2 = 0.65$; $P = 0.04$, Extended Data Fig. 5).

To calculate drought coupling, we first conducted a similar analysis to equation (1) by calculating the R^2 for each grid cell of the regression of VOD at midday versus VOD at night and daily VPD, although without the interaction term present in equation (3). The night-time (01:30) VOD was used because it is representative of root-zone soil moisture variation, given pre-dawn equilibrium between root-zone soil water potential and leaf water potential. This R^2 was used to measure drought coupling (coefficient of determination) at global scale. To quantify drought sensitivity, we used estimates of ecosystem 'anisohydricity' derived from the same VOD data. Anisohydricity is a measure of leaf water-potential response to soil drying and increases in VPD, as influenced by stomatal closure and xylem loss of conductivity. The anisohydricity was calculated as previously published[22] based on the slope of the relationship between midday (13:30 local time) overpass VOD and midnight (01:30 local time) overpass VOD, for which—at the latter time—leaf and soil water potential are assumed to be in equilibrium owing to night-time refilling. Higher values indicate more anisohydric or less drought-sensitive ecosystems, although prolonged droughts may still lead to a reduction in growth[46].

For the US-based analysis, we used the richness of dominant tree species from a previous study[47] to compare to hydraulic diversity of our eddy covariance sites. Globally, we used the previously published dataset[48] that provides spatial patterns of plant diversity based on species-area relationships and ground plots. We converted this dataset to a raster with 1×1 degree resolution and then used the previously published biome map[49] at the same resolution to determine the biome of each grid cell. We examined only grid cells that were dominated by one of six forest biomes (moist tropical forest, dry tropical forest, broadleaf temperate forest, coniferous temperate forest, boreal forest or Mediterranean-type forest/woodland). We further screened grid cells to remove cells with >25% human impact (typically croplands) based on a previous study[50].

We converted the vegetation water content variation metrics (correlation and sensitivity to drought variables) to the same resolution using bilinear interpolation. For each biome, we examined the relationship between species richness and water content variation using linear mixed-effects models to account for spatial auto-correlation. Per standard practice[51], we included the latitude and longitude coordinates of each grid cell in the regression and tested the following spatial correlation structures—linear, quadratic ratio, exponential, spherical and Gaussian—selecting the most likely and parsimonious model using $\Delta\mathrm{AIC} < -2$ or more. The quadratic correlation structure was most parsimonious for all biomes. We further checked that the effect of species richness was not due to patterns in biomass and/or productivity alone by including models that had both richness and the mean pixel VOD over the whole record, which is strongly related to aboveground biomass. Using stepwise model selection using AIC, the most parsimonious model included both mean VOD and species richness, indicating that richness is important beyond average biomass/productivity. Finally, we performed a similar random forest analysis and variable importance as conducted for the flux tower analysis and included species richness, mean VOD, canopy height[52] (to account for potential successional and/or land use effects) and biome.

Statistics. No statistical methods were used to predetermine sample size. All statistical analyses were performed in the R computing environment[53]. Model selection was performed using the stepAIC function in the MASS package[54]. Statistical assumptions of linear models were verified by examining residual and quantile plots of the models and transformations were applied as needed. All maps were generated using the raster[55] and rworldmap[56] packages and spatial autocorrelation was modelled using the gls function in the nlme package[57]. The generalized additive model in Extended Data Fig. 9 was performed using the gam function in the mgcv package[58]. Random forest analyses were performed using the randomForest and ranger packages.

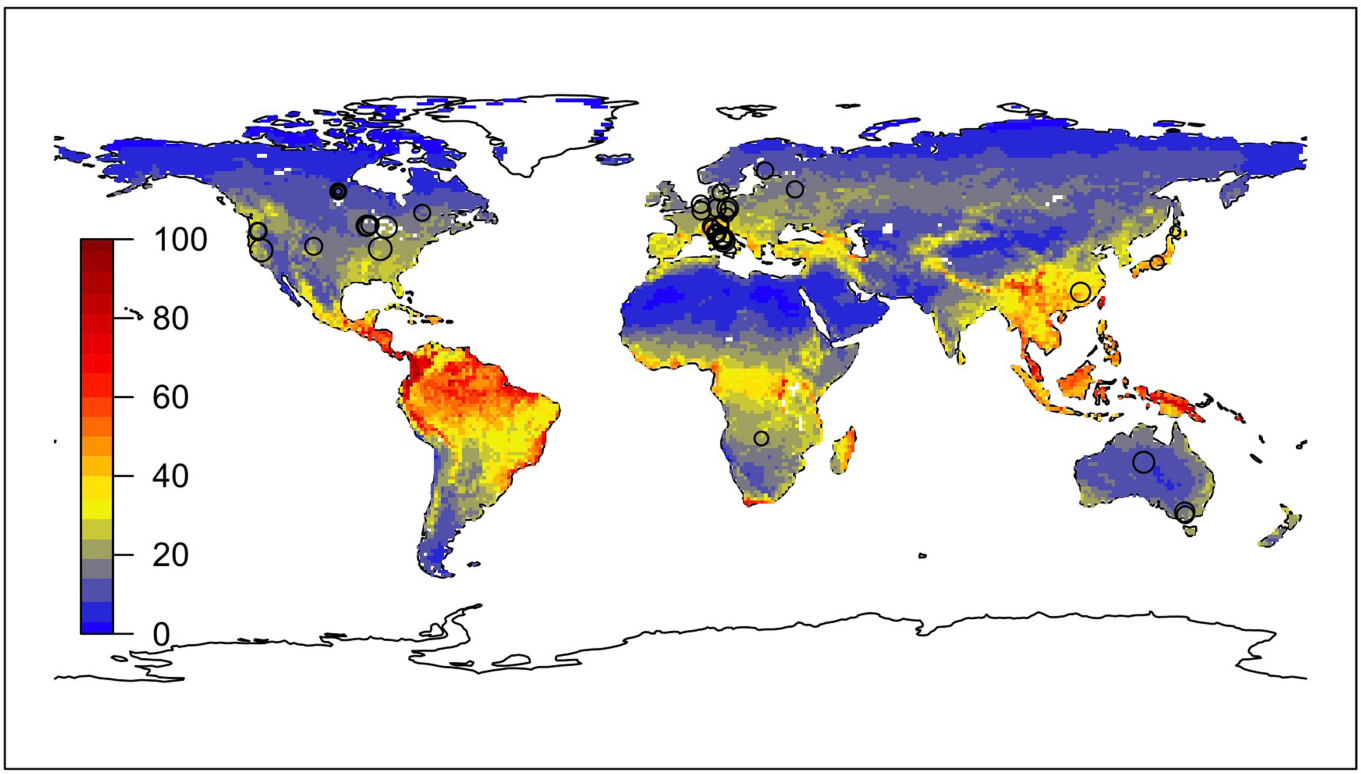
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Eddy flux data are available at http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/; community trait data are available at http://www.anderegglab.net/data/trait-data/; detailed trait data are available in Extended Data Figs. 1–10 and at https://datadryad.org//handle/10255/dryad.235, https://datadryad.org//handle/10255/dryad.80340 and from a previous publication[18].

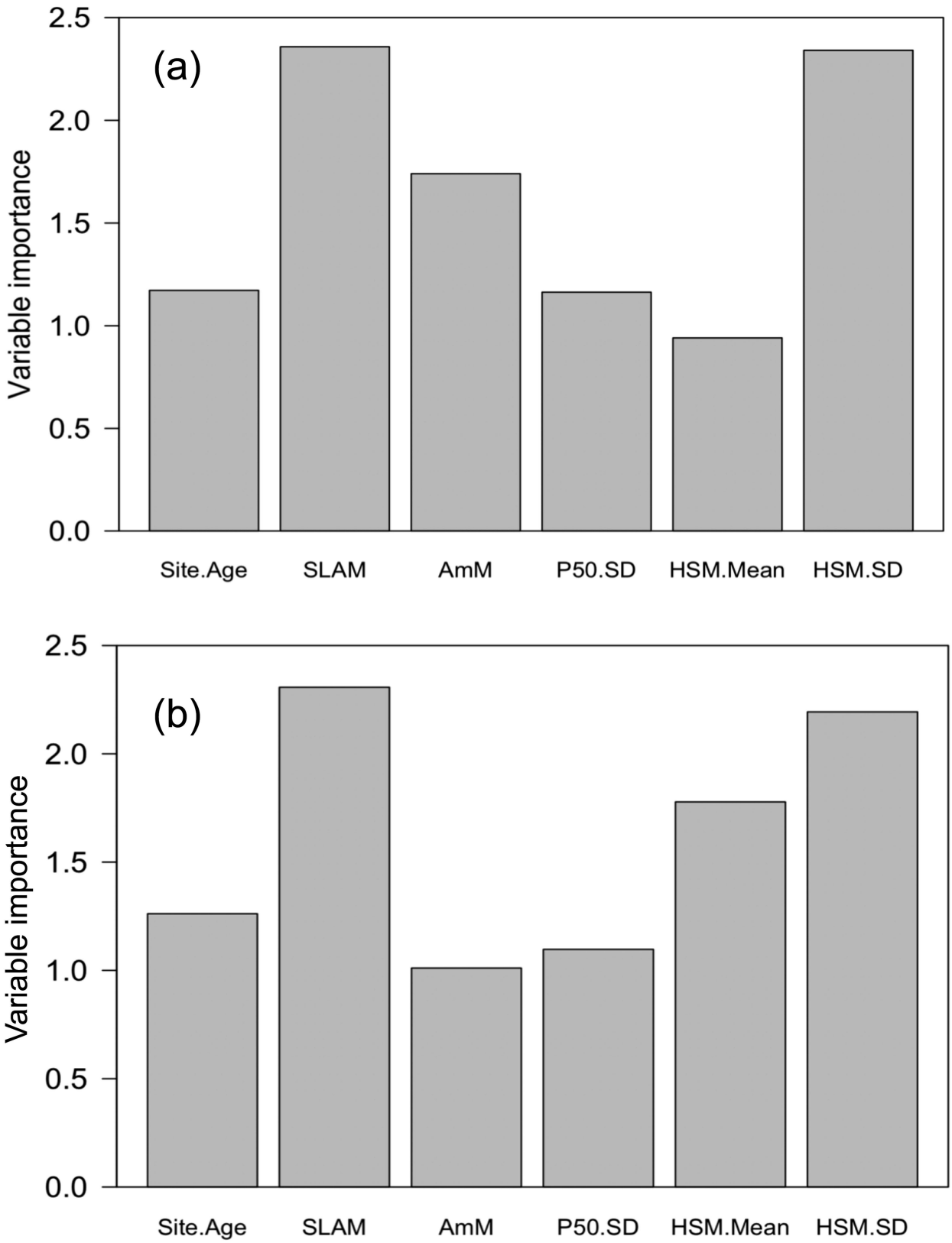
31. Pastorello, G. Z. et al. The FLUXNET2015 dataset: The longest record of global carbon, water, and energy fluxes is updated. *Eos* **98**, (2017).
32. Clogg, C. C., Petkova, E. & Haritou, A. Statistical methods for comparing regression coefficients between models. *Am. J. Sociol.* **100**, 1261–1293 (1995).
33. Zanne, A. E. et al. Global wood density database. *Dryad* https://doi.org/10.5061/dryad.234/1 (2009).
34. Chave, J. et al. Towards a worldwide wood economics spectrum. *Ecol. Lett.* **12**, 351–366 (2009).
35. Maire, V. et al. Global effects of soil and climate on leaf photosynthetic traits and rates. *Glob. Ecol. Biogeogr.* **24**, 706–717 (2015).
36. Gleason, S. M. et al. Weak tradeoff between xylem safety and xylem-specific hydraulic efficiency across the world's woody plant species. *New Phytol.* **209**, 123–136 (2016).
37. Anderegg, L. D. L., Anderegg, W. R. L., Abatzoglou, J., Hausladen, A. M. & Berry, J. A. Drought characteristics' role in widespread aspen forest mortality across Colorado, USA. *Glob. Change Biol.* **19**, 1526–1537 (2013).
38. Burnham, K. P. & Anderson, D. R. Multimodel inference understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304 (2004).
39. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
40. Jackson, T. J. & Schmugge, T. J. Vegetation effects on the microwave emission of soils. *Remote Sens. Environ.* **36**, 203–212 (1991).
41. Momen, M. et al. Interacting effects of leaf water potential and biomass on vegetation optical depth. *J. Geophys. Res. Biogeosci.* **122**, 3031–3046 (2017).
42. Du, J., Kimball, J. S. & Jones, L. A. Passive microwave remote sensing of soil moisture based on dynamic vegetation scattering properties for AMSR-E. *IEEE Trans. Geosci. Remote Sens.* **54**, 597–608 (2016).
43. Du, J. et al. A global satellite environmental data record derived from AMSR-E and AMSR2 microwave earth observations. *Earth Syst. Sci. Data* **9**, 791–808 (2017).
44. Jones, L. A. et al. Satellite microwave remote sensing of daily land surface air temperature minima and maxima from AMSR-E. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **3**, 111–123 (2010).
45. Huffman, G. J. et al. Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeorol.* **2**, 36–50 (2001).
46. Konings, A. G., Williams, A. P. & Gentine, P. Sensitivity of grassland productivity to aridity controlled by stomatal and xylem regulation. *Nat. Geosci.* **10**, 284–288 (2017).
47. Jenkins, C. N., Van Houtan, K. S., Pimm, S. L. & Sexton, J. O. US protected lands mismatch biodiversity priorities. *Proc. Natl Acad. Sci. USA* **112**, 5081–5086 (2015).
48. Ellis, E. C., Antill, E. C. & Kreft, H. All is not loss: plant biodiversity in the Anthropocene. *PLoS ONE* **7**, e30535 (2012).
49. Olson, D. M. et al. Terrestrial ecoregions of the world: a new map of life on earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience* **51**, 933–938 (2001).
50. Sanderson, E. W. et al. The human footprint and the last of the wild. *Bioscience* **52**, 891–904 (2002).
51. Dormann, C. F. et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609–628 (2007).
52. Simard, M., Pinto, N., Fisher, J. B. & Baccini, A. Mapping forest canopy height globally with spaceborne lidar. *J. Geophys. Res.* **116**, G04021 (2011).
53. R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, 2012).
54. Ripley, B. et al. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-50 https://cran.r-project.org/web/packages/MASS/index.html (2013).
55. Hijmans, R. J. & van Etten, J. *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7 https://cran.r-project.org/web/packages/raster/index.html (2014).
56. South, A. *rworldmap: Mapping Global Data*. R package version 1.3-6 https://cran.r-project.org/web/packages/rworldmap/index.html (2011).
57. Bates, D. et al. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. R package version 1.1-18-1 https://cran.r-project.org/web/packages/lme4/index.html (2014).
58. Wood, S. N. mgcv: GAMs and generalized ridge regression for R. *R News* **1** (2), 20–25 (2001).

© 2018 Springer Nature Limited. All rights reserved.



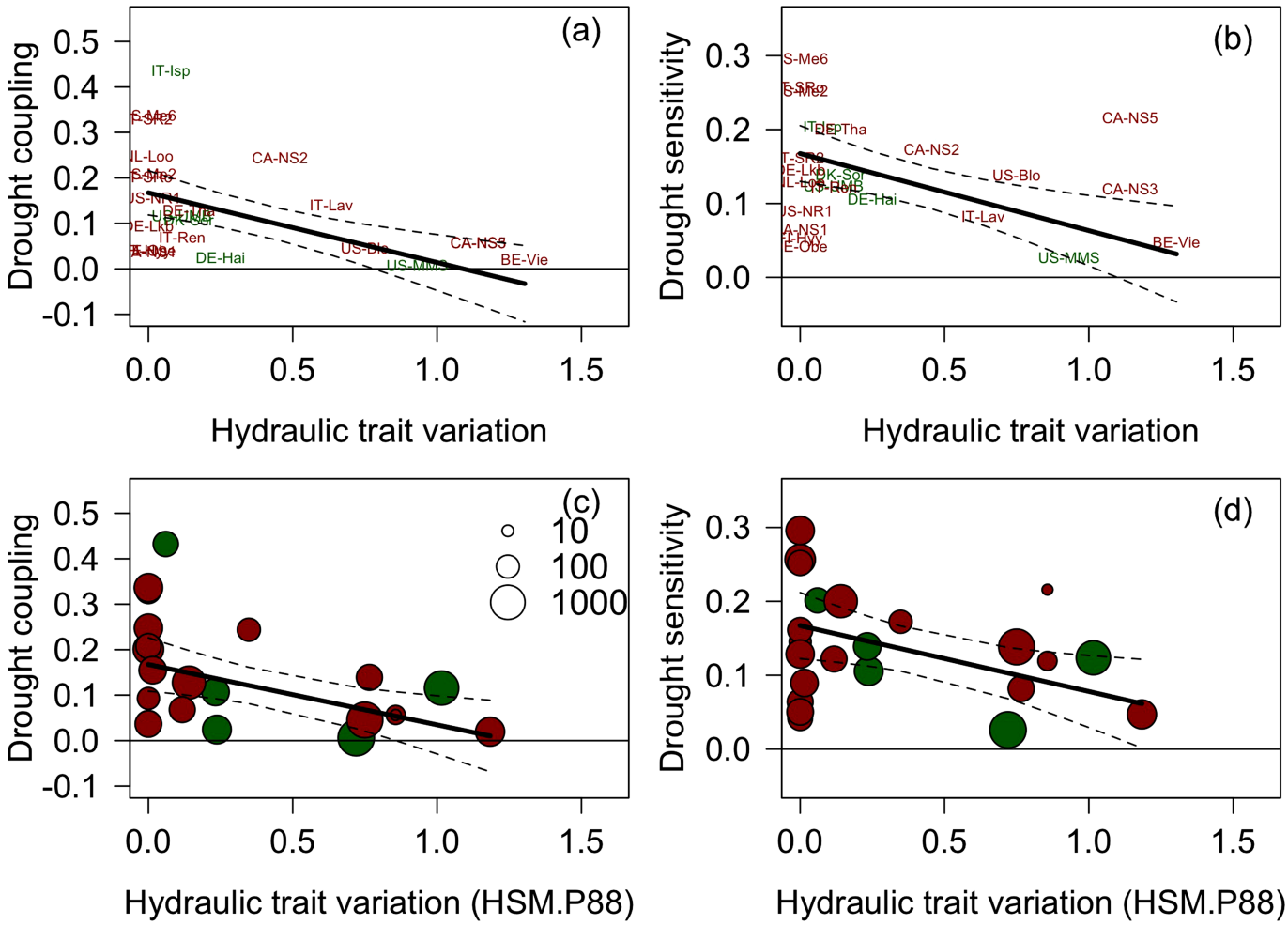
Extended Data Fig. 1 | Map of the included eddy covariance flux sites overlaid on species richness. Species richness is shown by different colours (data from http://ecotope.org/anthromes/biodiversity/plants/data/). Size of the circle is representative of the sample size of included days ranging from 10 (smallest circles) to 1,057 (largest circles) samples.

© 2018 Springer Nature Limited. All rights reserved.



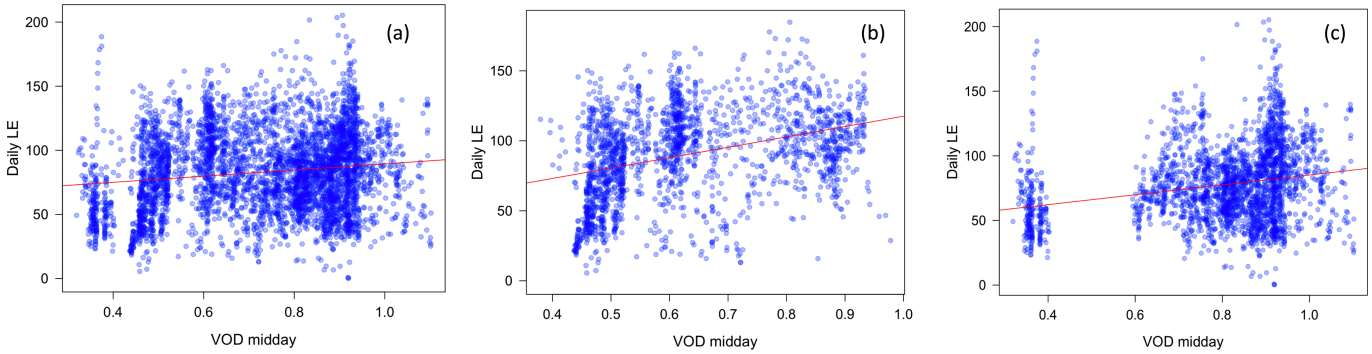
Extended Data Fig. 2 | Variable importance analysis for traits at flux sites. a, **b**, Variable importance (total decrease in node impurities) results from the machine-learning algorithm, random forests, for each variable for drought sensitivity (**a**) and drought coupling (**b**) metrics. Traits include SLA, (A_{max}), P_{50} and HSM. The suffix 'm' indicates the community-weighted mean; the suffix 'SD' indicates the community-weighted standard deviation. See Supplementary Table 3 for sample sizes.

© 2018 Springer Nature Limited. All rights reserved.



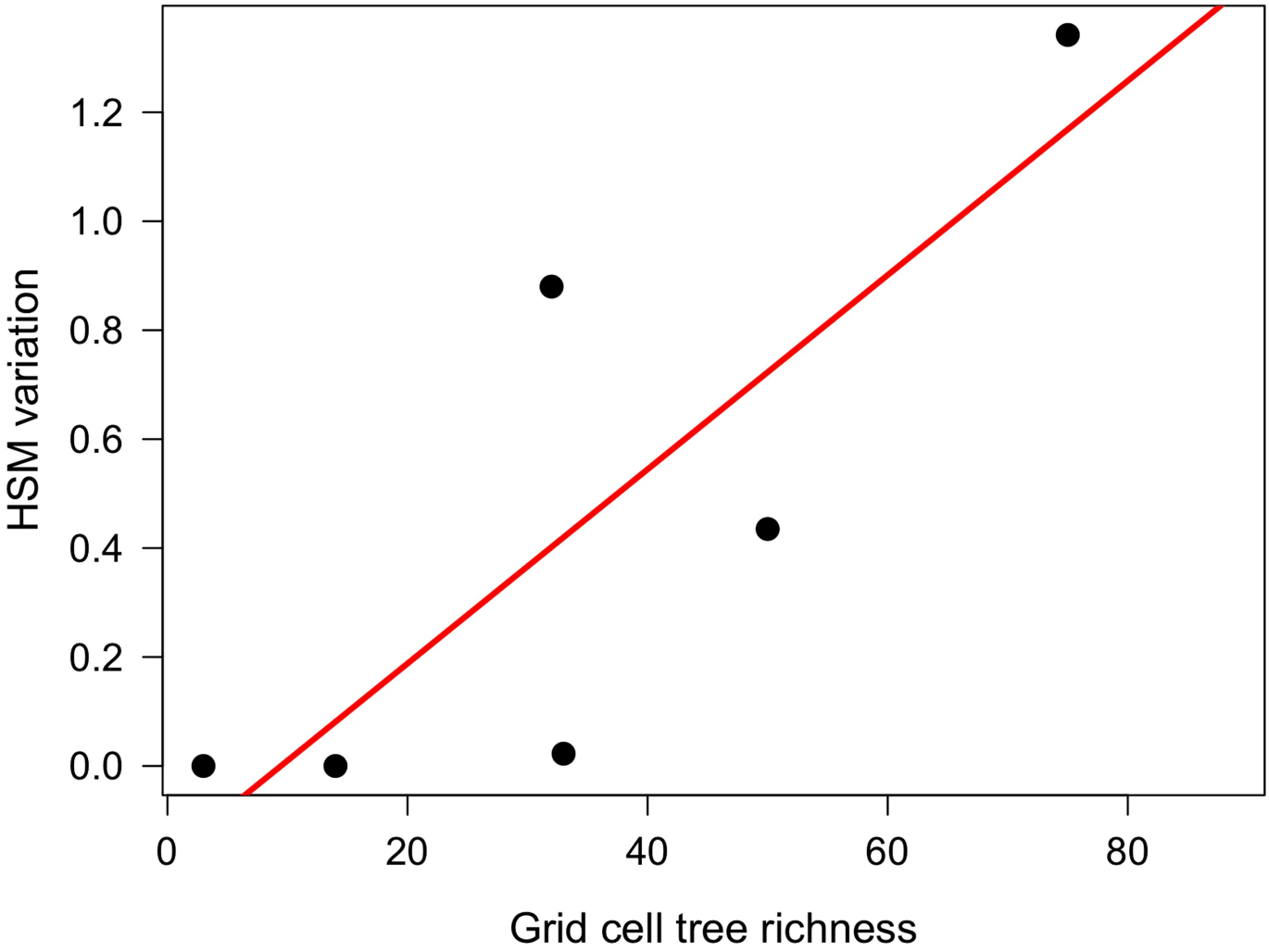
Extended Data Fig. 3 | Increased hydraulic variation buffers ecosystem drought responses. a, **c**, Drought coupling is expressed as the percentage of explained variation (R^2). **b**, **d**, Drought sensitivity is shown as the summed absolute values of standardized coefficients for drought variables that are regressed against latent energy (LE) exchange. Regression: LE ≈ VPD + SM + VPD × SM. **a**, **b**, Panels are identical to Fig. 2 but with site identifications shown. Hydraulic variation is expressed as the community-weighted standard deviation in the hydraulic safety margin of species. **c**, **d**, The hydraulic safety margin was calculated from the 50% loss of hydraulic conductivity in gymnosperms and 88% loss of hydraulic conductivity in angiosperms. Colours indicate biomes of deciduous broadleaf (green) and needleleaf (red) forests. The size of the dot indicates the number of days included for each flux site. The solid black line is the best fit of the ordinary least-squares linear regression (**c**, $P = 0.008$; **d**, $P = 0.01$) and dashed lines are the 95% confidence interval of the regression fit.

© 2018 Springer Nature Limited. All rights reserved.



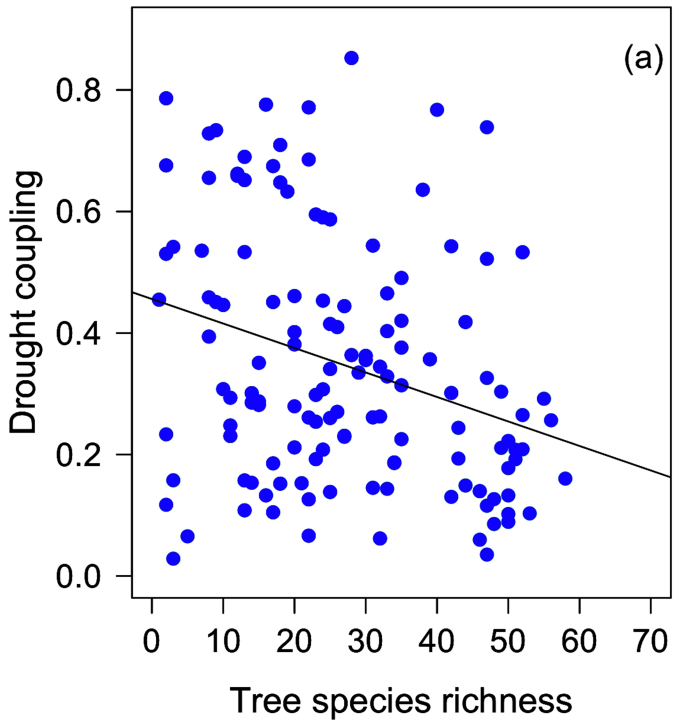
Extended Data Fig. 4 | Satellite comparisons to flux towers. **a**–**c**, Relationship between daily total latent energy (LE) exchange measured via eddy covariance and midday canopy water content from remote-sensing of VOD for global forest sites (**a**; $n = 4{,}525$ grid cells), broadleaf forest sites (**b**; $n = 1{,}915$ grid cells) and evergreen forest sites (**c**; $n = 2{,}610$ grid cells). Red lines indicate best fits for ordinary least-squares regressions. Note that the canopy water content at each pixel integrates a spatial area that is two orders of magnitude greater than the eddy covariance sites.

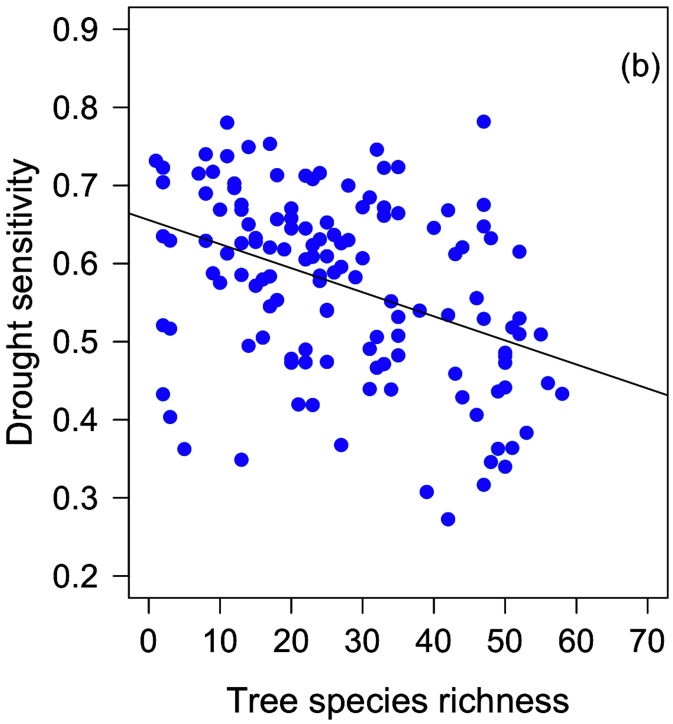
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Hydraulic trait variation compared to species richness. Comparison of dominant tree species richness from gridded data of US forests against the hydraulic diversity—the standard deviation in HSM—at six eddy covariance sites in the United States that have adequate trait data. The red line indicates the best fit of the ordinary least-squares linear regression ($n = 6$ sites; $R^2 = 0.65$; $P = 0.04$).

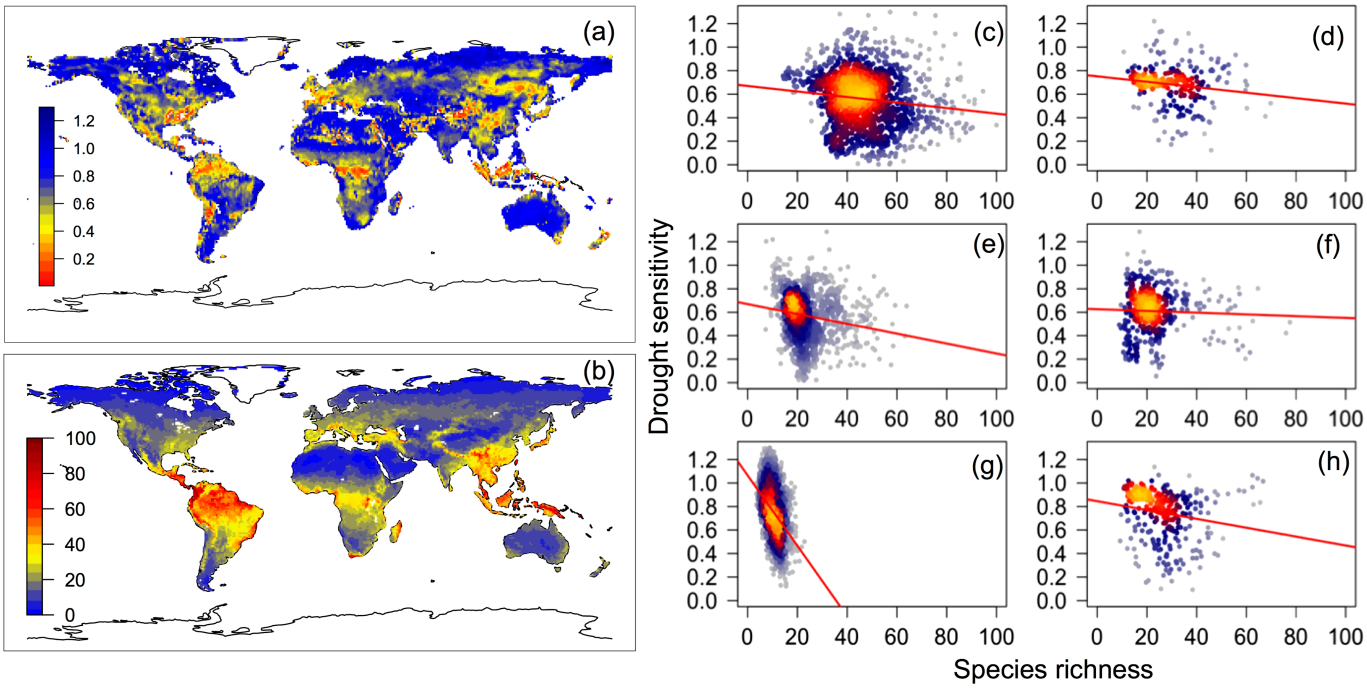
© 2018 Springer Nature Limited. All rights reserved.





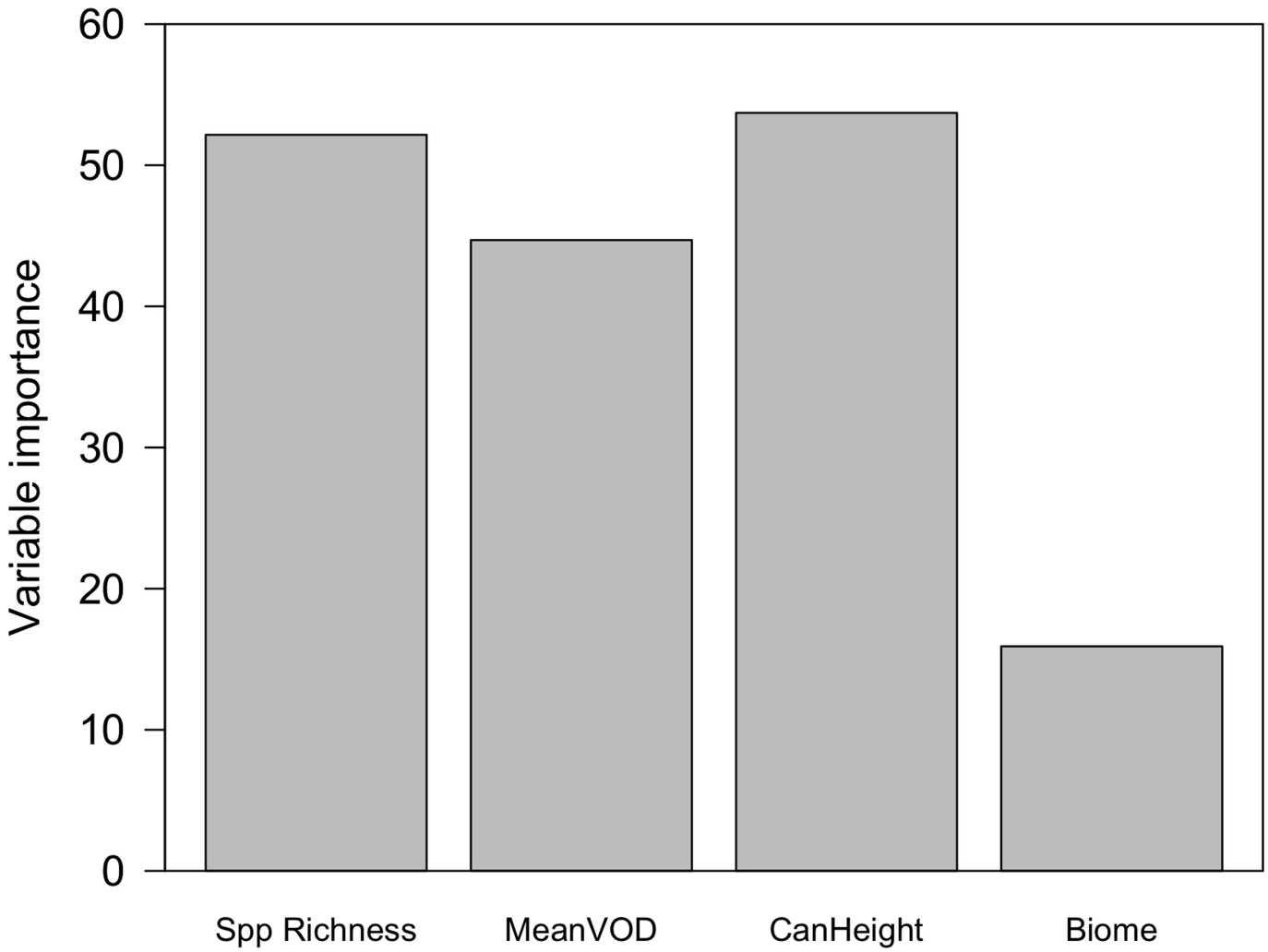
Extended Data Fig. 6 | Higher species diversity is associated with more buffered drought responses in US forests. a, **b**, Drought coupling (**a**; $n = 163$ grid cells) and drought sensitivity (**b**; $n = 163$ grid cells) from variation in remotely sensed canopy vegetation water content compared to tree species richness in the United States. Black line is the best fit of the ordinary least-squares linear regressions.

© 2018 Springer Nature Limited. All rights reserved.



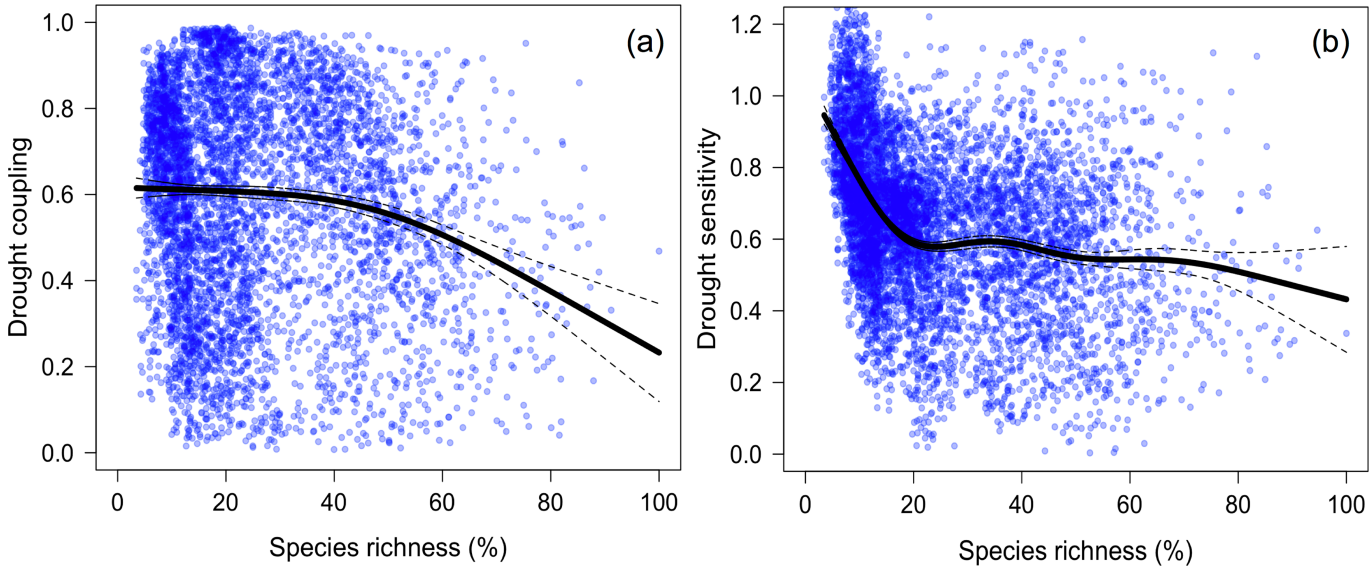
Extended Data Fig. 7 | Higher species diversity is associated with more-buffered drought responses in forests globally. a, Drought sensitivity as the slope (β) in an ordinary least-squares linear regression of an index of the variation in aboveground plant water content at midday compared to night (regression: $VOD_{midday} = \beta \times VOD_{night} + \varepsilon$). **b**, Native plant species richness (percentage of maximum; data from http://ecotope.org/anthromes/biodiversity/plants/data/). **c**–**h**, Ordinary least-squares linear regressions between these two variables for six major biomes. **c**, Tropical and subtropical moist broadleaf forests ($n = 1{,}380$ grid cells). **d**, Tropical and subtropical dry broadleaf forests ($n = 241$ grid cells). **e**, Temperate broadleaf and mixed forests ($n = 1{,}289$ grid cells). **f**, Temperate coniferous forests ($n = 318$ grid cells). **g**, Boreal forests ($n = 1{,}784$ grid cells). **h**, Mediterranean-type forests, woodlands and shrub ($n = 319$ grid cells). Each point represents an individual grid cell from the map and colours that are more red indicate a higher density of points. Red lines show the best fit of ordinary least-squares linear regression lines.

© 2018 Springer Nature Limited. All rights reserved.



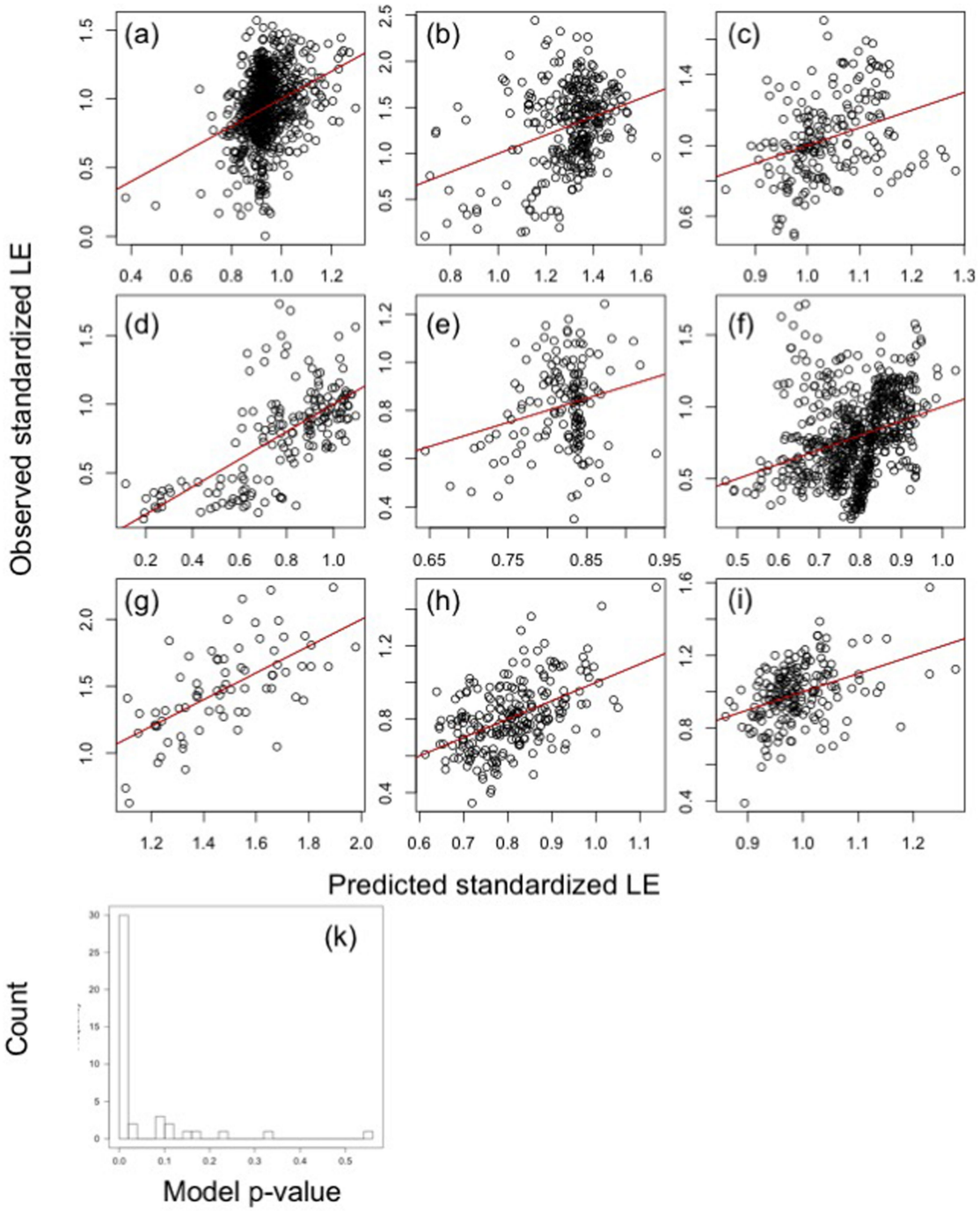
Extended Data Fig. 8 | Analyses of the importance of variables using satellite data. The importance of the variables (total decrease in node impurities) results obtained using the machine learning algorithm, random forests, for each variable of the drought coupling metric ($n = 6{,}698$ grid cells). 'CanHeight', lidar-derived canopy height.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Ecosystem sensitivity to drought saturates with species richness (the percentage of maximum) across forests globally. **a**, Drought coupling is expressed as the explained variation (R^2) of midday aboveground plant water content in forest ecosystems regressed against drought variables using ordinary least-squares linear regression. **b**, Drought sensitivity is expressed as the regression coefficient of midday aboveground plant water content regressed against a metric of soil water stress using ordinary least-squares linear regression. $n = 6{,}698$ grid cells. The black line shows the best fit generalized additive model and dashed lines are the 99% confidence interval.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Performance of the multivariate drought regression model. Nine randomly selected sites of observed latent energy (LE) fluxes versus predicted fluxes from the multiple regression model based on VPD, soil moisture and their interaction. Red lines are the ordinary least-squares best-fit regression line. Sites are as follows (abbreviations can be found in Supplementary Table 1). **a**, US-UMB ($R^2 = 0.12$, $P < 0.0001$). **b**, US-WCr ($R^2 = 0.13$, $P < 0.0001$). **c**, DK-Sor ($R^2 = 0.11$, $P < 0.0001$). **d**, IT-Ca1 ($R^2 = 0.44$, $P < 0.0001$). **e**, IT-PT1 ($R^2 = 0.06$, $P = 0.002$). **f**, IT-Ro1 ($R^2 = 0.10$, $P < 0.0001$). **g**, JP-SMF ($R^2 = 0.40$, $P < 0.0001$). **h**, NL-Loo ($R^2 = 0.25$, $P < 0.0001$). **i**, US-NWR ($R^2 = 0.15$, $P < 0.0001$). **k**, Performance of the drought multivariate ordinary least-squares linear regression model ($LE = f(VPD, SM, VPD \times SM)$) across eddy covariance sites shown in a histogram of site-level model P values from those regressions.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0527-y

Sensation, movement and learning in the absence of barrel cortex

Y. Kate Hong[1], Clay O. Lacefield[1], Chris C. Rodgers[1] & Randy M. Bruno[1]*

For many of our senses, the role of the cerebral cortex in detecting stimuli is controversial[1–17]. Here we examine the effects of both acute and chronic inactivation of the primary somatosensory cortex in mice trained to move their large facial whiskers to detect an object by touch and respond with a lever to obtain a water reward. Using transgenic mice, we expressed inhibitory opsins in excitatory cortical neurons. Transient optogenetic inactivation of the primary somatosensory cortex, as well as permanent lesions, initially produced both movement and sensory deficits that impaired detection behaviour, demonstrating the link between sensory and motor systems during active sensing. Unexpectedly, lesioned mice had recovered full behavioural capabilities by the subsequent session. This rapid recovery was experience-dependent, and early re-exposure to the task after lesioning facilitated recovery. Furthermore, ablation of the primary somatosensory cortex before learning did not affect task acquisition. This combined optogenetic and lesion approach suggests that manipulations of the sensory cortex may be only temporarily disruptive to other brain structures that are themselves capable of coordinating multiple, arbitrary movements with sensation. Thus, the somatosensory cortex may be dispensable for active detection of objects in the environment.

Sensory detection tasks have become a staple for probing cortical circuitry during behaviour, but the role of the primary sensory cortex in visual[1–3], auditory[4–6], gustatory[7,8] and somatosensory behaviours[9–17] remains unclear. Whether a brain structure is necessary for a behaviour is typically assessed by inactivation or ablation. Ablation experiments may underestimate behavioural deficits, owing to the long recovery periods used (more than one week), during which compensatory relearning or rewiring can occur. Transient optogenetic or pharmacological manipulations often yield stronger deficits and are currently preferred, being thought to reveal an area's normal function before compensation. However, the sudden loss of a silenced area may disrupt downstream areas that are vital for behaviour, a phenomenon

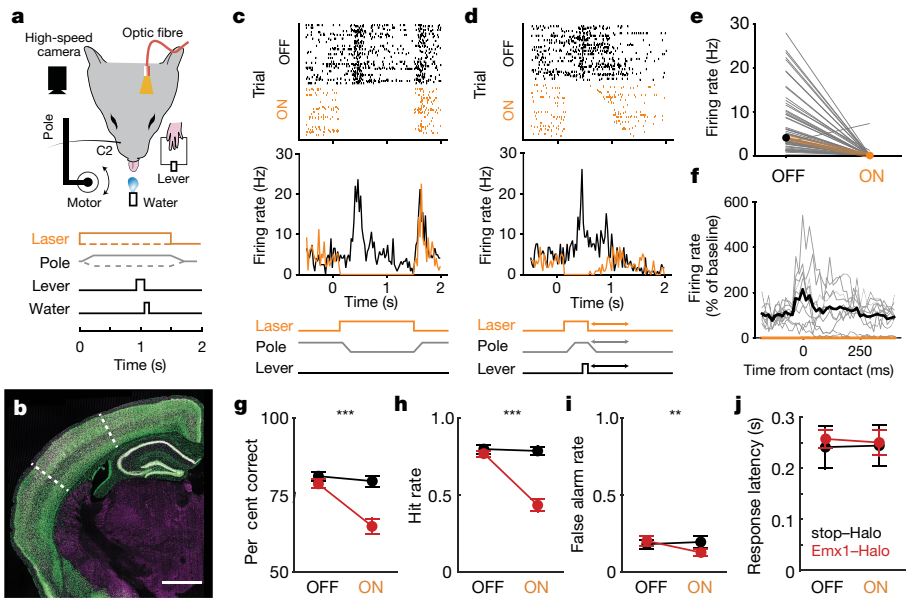


Fig. 1 | Transient inactivation of barrel cortex impairs whisker-mediated detection. a, Head-fixed detection task. A high-speed camera imaged whiskers. Pole movement (GO or NOGO) and laser (ON or OFF) were randomized across trials. **b**, Coronal section of Emx1–GFP mouse brain. White dotted lines, barrel cortex boundaries; green, GFP; magenta, NeuN. Scale bar, 1 mm. **c**, Cortical array recordings during detection task and photoinactivation. Example rasters (top) and peri-stimulus time histograms (middle) for a single unit when no response was made (miss or correct reject) during an example session. **d**, Same neuron for trials with lever response (hit or false alarm). The laser turns off when the mouse responds, ending the trial. Trials sorted by response time, which varied (arrows in schematic at bottom). **e**, Effect of laser on neuronal spiking ($n = 62$ putative excitatory neurons, 8 sessions, 3 mice; mean ± s.d. 7.02 ± 6.81 and 0.16 ± 0.94 Hz for laser OFF versus ON, respectively). **f**, Average spiking activity aligned to first whisker contact of trial during an example session ($n = 10$ neurons). Contact times are defined as the local maxima, rather than onset, of curvature change (Extended Data Fig. 3; see Methods). Firing rates were normalized to mean rate during 100 ms before contact during laser-OFF trials. Thin lines, individual cells; thick lines, means of OFF (black) and ON (orange). **g**–**i**, Behavioural performance for laser OFF versus ON trials. **g**, Per cent correct trials. **h**, Hit rate. **i**, False alarm rate. **j**, Response latency for hit trials. Emx1–Halo ($n = 10$ mice, red), negative control ($n = 7$ cre-negative, stop–Halo mice, black). Data shown as mean ± s.e.m. $^{**}P < 0.01$, $^{***}P < 0.001$.

[1]Department of Neuroscience, Mortimer Zuckerman Mind Brain Behavior Institute and Kavli Institute for Brain Science, Columbia University, New York, NY, USA. *e-mail: randybruno@columbia.edu

© 2018 Springer Nature Limited. All rights reserved.

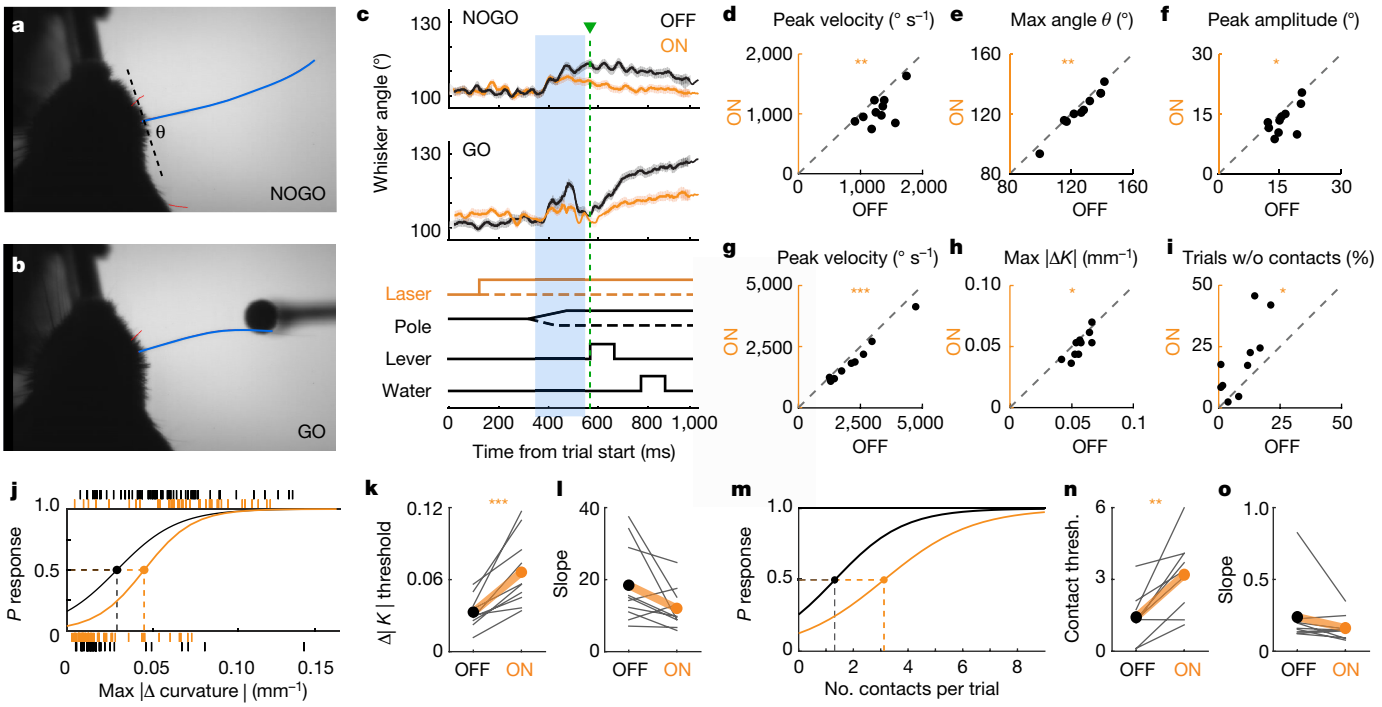


Fig. 2 | Transient optogenetic inactivation of barrel cortex alters whisking kinematics and sensory threshold. a, **b**, High-speed video frame depicting traced C2 whisker during NOGO trial (**a**, pole moves away) and GO trial (**b**, pole within whisker reach). Whisker position was measured as its angle (θ) relative to the face. The whisker bends upon contacting the pole, changing whisker curvature. **c**, Average whisker angle for NOGO and GO trials for an example session. A 200-ms window (blue shaded area), from when the pole was within reach and before the response, was analysed. Green, average response time. **d**–**f**, Whisking kinematics for each mouse during NOGO trials. **d**, Peak angular velocity of whisker protraction; **e**, maximum whisker angle; **f**, mean peak whisking amplitude. **g**–**i**, For GO trials: **g**, peak angular protraction velocity; **h**, average maximum change in curvature (ΔK); **i**, per cent of trials without any contacts. **j**, Logistic regression of response probability given maximum change in curvature for an example session. Tick marks indicate responses (0 for no response; 1 for lever response) on individual trials. Detection threshold was defined as the value at which response probability is 0.5. **k**, Detection threshold for maximum ΔK of each mouse. Thick orange line, means. **l**, Slope (sensitivity). **m**, Logistic regression of response probability against number of contacts per trial for an example session. **n**, **o**, Detection threshold (**n**) for number of contacts and slope (**o**). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $n = 10$ mice.

known as diaschisis. Recent studies in the motor system have underscored how the off-target effects of transient inactivation can lead to false conclusions[18]. To address these disparate outcomes, we compare transient and chronic manipulations of the barrel cortex subdivision of the primary somatosensory cortex (S1) during a simple detection task.

Water-restricted mice were trained in the dark to perform a GO/NOGO sensory detection task with their C2 whisker (Fig. 1a). Mice self-initiated trials by holding down a lever. On GO trials, a pole moved within reach of the whisker when protracted. Mice had to release the lever when they detected the pole to obtain a water reward (hit). On NOGO trials, the pole moved away from the mouse. Incorrect responses to NOGO trials (false alarms) were punished with a timeout. Misses and correct rejects were neither rewarded nor punished.

Conventionally, cortex is optogenetically silenced by activating inhibitory cells with channelrhodopsin (ChR), but this may inadvertently stimulate long-range inhibitory connections. We therefore developed an approach to directly silence excitatory cells by stably expressing halorhodopsin (Halo) in cortical excitatory neurons. Emx1-IRES-cre mice express Cre recombinase in excitatory cortical neurons while excluding subcortical structures (Fig. 1b). Halorhodopsin can be targeted to these neurons by crossing Emx1-IRES-cre with stop-Halo reporter mice (Emx1–Halo). Optogenetic silencing (by shining a laser onto these neurons during behaviour) was highly efficacious, blocking 95 ± 4% (mean ± s.e.m.) of spikes in putative excitatory cells (Fig. 1c–e, Extended Data Fig. 1j, k) including during whisker contacts, which normally strongly activate barrel cortex (Fig. 1f). Optogenetics efficiently silenced spontaneous and sensory-evoked activity in neurons across all cortical layers within a 1-mm radius, encompassing nearly all of the barrel columns that represent the large facial whiskers (Extended Data Fig. 1a–i).

Inactivation of barrel cortex significantly reduced overall performance ($P = 3.6 \times 10^{-4}$; $n = 10$ mice; Fig. 1g, Extended Data Fig. 2a), but performance remained significantly above chance (50%; $P = 1.1 \times 10^{-8}$, one-sample t-test). Despite the fact that mice made fewer responses to both GO and NOGO trials (Fig. 1h, i), response time was unaffected by the laser (Fig. 1j), demonstrating that the mice could still manoeuvre the lever. Control mice lacking Cre were behaviourally unaffected by the laser ($n = 7$ stop–Halo mice; Fig. 1g–j). Conventional photoinhibition by activating parvalbumin (PV)-positive inhibitory cells ($n = 5$ PV–ChR mice) yielded similar results to Emx1–Halo (Extended Data Fig. 2a). Thus, transient silencing of barrel cortex impaired detection.

Touch is an active process, during which subjects adjust their movements as they contact objects in their environment[19–21]. Even small changes in whisking could alter perception[19]. Although activation of barrel cortex can trigger whisker movements[16,22], the effects of inactivation are less well understood[12,23]. We tracked whisking using high-speed videography (Fig. 2a, b). During NOGO trials, in which no contacts were possible, barrel cortex inactivation slightly but significantly decreased protraction velocity, whisker angle and peak amplitude (Fig. 2c–f, $P = 6.5 \times 10^{-3}$, 1.4×10^{-3} and 2.5×10^{-2}, for d, e and f, respectively). Similarly, during GO trials, when the pole was present inactivation of barrel cortex decreased peak protraction velocity (Fig. 2g). We found no significant changes in whisking setpoint or frequency. We also assessed changes in whisker curvature, a proxy for contact force[24] (Extended Data Fig. 3). Small changes in whisker movement had a large effect on whisker contacts, resulting in a reduction in force (Fig. 2h) and an increase in the number of trials without contacts (Fig. 2i). Thus, silencing of the sensory cortex reduced the vigour of whisker movement.

© 2018 Springer Nature Limited. All rights reserved.

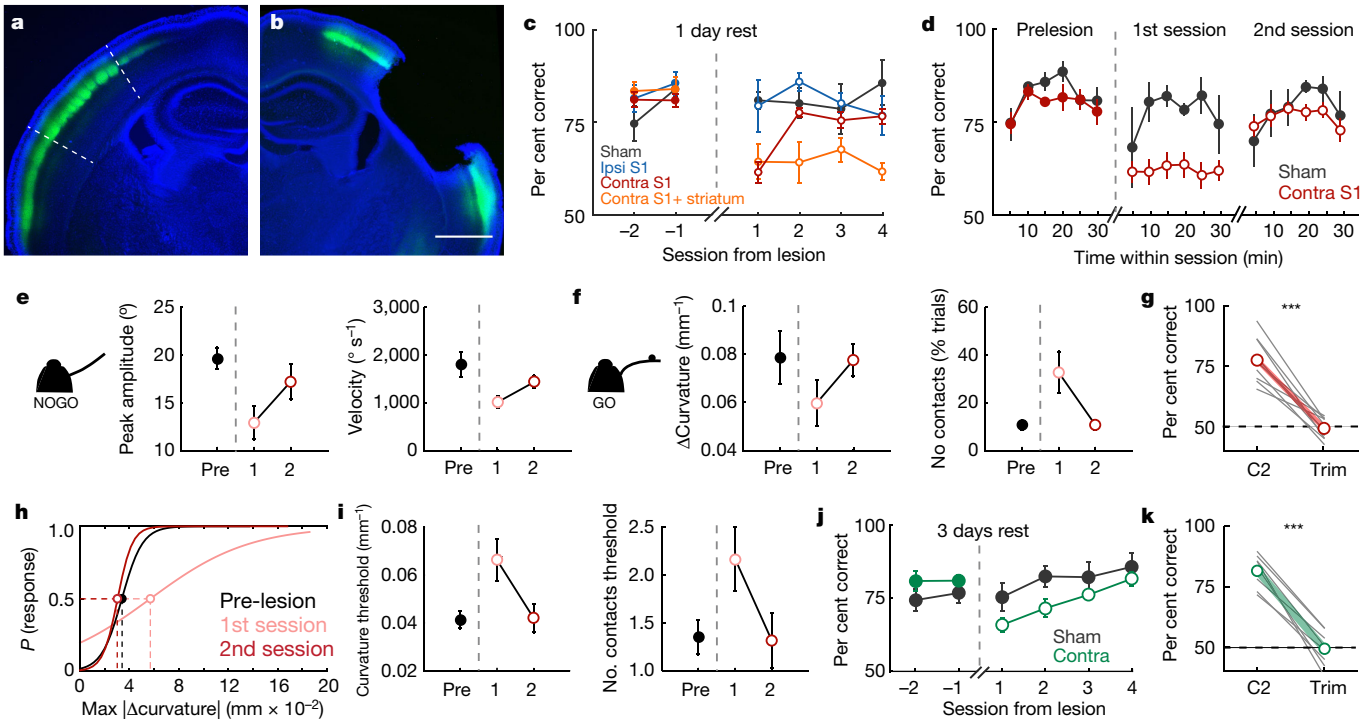


Fig. 3 | Behavioural performance recovers rapidly after barrel cortex lesions. a, Coronal section of L4-labelled mouse brain (white dashed lines, barrel cortex boundaries; blue, DAPI; green, eYFP) in unlesioned hemisphere. **b**, Lesioned hemisphere shows complete removal of barrel cortex. **c**, Behavioural performance before and after S1 lesions (sham $n=5$, ipsilateral $n=4$, contralateral $n=11$, contralateral with striatal lesion $n=8$). **d**, Behavioural performance of contralateral S1 lesioned mice recovers abruptly between first and second sessions after lesioning. **e**, **f**, Whisking kinematics during NOGO trials (**e**) and GO trials of mice (**f**) are altered on first post-lesion session but return to normal by second post-lesion session ($n=10$ contralaterally lesioned mice). **g**, Post-lesion trimming confirms that task remained whisker-dependent. Dashed line indicates chance performance. **h**, Example logistic regression of response probability given whisker curvature. Sensory detection threshold for whisker curvature increases on first session post-lesion (pink) and returns to pre-lesion levels (black) by second session (red). **i**, Average detection thresholds for whisker curvature and number of contacts before lesion and for second session after lesion ($n=10$). **j**, Mice with three days of rest after lesion still had impaired performance on the first post-lesion session but subsequently recovered ($n=8$). **k**, Behavioural performance for three-day rest group remains whisker-dependent. $^{***}P<0.001$.

We investigated whether behavioural impairment was simply due to altered whisking or whether there was an accompanying sensory deficit: whether, for any given stimulus strength, transient inactivation of the sensory cortex decreased the probability of response. Cortical silencing significantly increased detection threshold (0.5 response probability) for curvature (Fig. 2j, k; $P=9.7\times10^{-4}$) and number of contacts (Fig. 2m, n; $P=3.3\times10^{-3}$), but not sensitivity (Fig. 2l, o). We observed similar motor and sensory deficits in PV–ChR mice (Extended Data Fig. 2b, c). Thus, transient optogenetic manipulations impair behaviour by both increasing sensory threshold and decreasing whisker movement.

Increased sensory threshold is distinct from an absolute inability to detect stimuli. The observed threshold shift could reflect incomplete inactivation, as a few renegade spikes may suffice for detection[25]. However, residual spiking during optogenetic silencing did not correlate with behavioural outcome (Extended Data Fig. 1l, m). To ensure complete inactivation, we removed contralateral barrel cortex by aspiration ($n=11$) (Fig. 3a, b, Extended Data Fig. 4). Consistent with optogenetic results, behaviour was impaired one day after lesioning contralateral barrel cortex (Fig. 3c, red), but not in sham-operated controls ($n=4$, black) or when ipsilateral barrel cortex was lesioned ($n=4$, blue). Again, impairment was only partial, and behaviour remained significantly above chance levels ($P<10^{-6}$, one-sample t-test).

Unexpectedly, by the second session after lesioning, behaviour had recovered fully to pre-lesion levels (Fig. 3c). Mice recovered whether lesions encompassed only barrel cortex or additionally included the secondary somatosensory cortex (Extended Data Fig. 4). There was no evidence of gradual relearning within sessions; rather, performance abruptly recovered between the first and second post-lesion sessions (Fig. 3d), suggesting that recovery was unlikely to result from previously uninvolved circuits learning the task anew. Furthermore, subsequent lesioning of ipsilateral barrel cortex did not perturb performance, as these bilaterally lesioned mice performed similarly to sham and ipsilateral-only lesioned mice (Extended Data Fig. 5). Notably, additional damage to the dorsolateral striatum prevented behavioural recovery (Fig. 3c, $n=8$, orange; Extended Data Fig. 6), suggesting that the striatum has an important role in detection behaviour.

Consistent with optogenetic results, whisker movement and contacts were decreased during the first session after lesioning (Fig. 3e, f, pre versus 1). However, whisking kinematics for the recovered, second session never exceeded pre-lesion levels (Fig. 3e, f; pre versus 2), indicating that mice did not compensate for impaired sensation with greater contact force or frequency. Similarly, sensory thresholds pre-lesion and on the second post-lesion session did not differ significantly in contact force or number (Fig. 3h, i). Thus, after only a temporary impairment, both motor and sensory abilities returned to pre-lesion levels along with behavioural performance.

Recent studies have suggested that homeostasis may spontaneously restore activity in connected structures within 24–48 h, similar to the time frame seen here[18,26]. To test whether behavioural recovery was spontaneous or required re-exposure to the task, we gave another group of mice three days between lesion and retesting ($n=8$). This group showed similar impairment on the first post-lesion session but also recovered (Fig. 3j), albeit more gradually, indicating that task re-exposure—rather than simply the passage of time—triggers recovery. Removing the C2 whisker reduced the performance of both groups (1 day rest and 3 days rest) to chance, confirming that lesions did not induce mice to switch from whisker-mediated touch to other sensory modalities (Fig. 3g, k).

© 2018 Springer Nature Limited. All rights reserved.

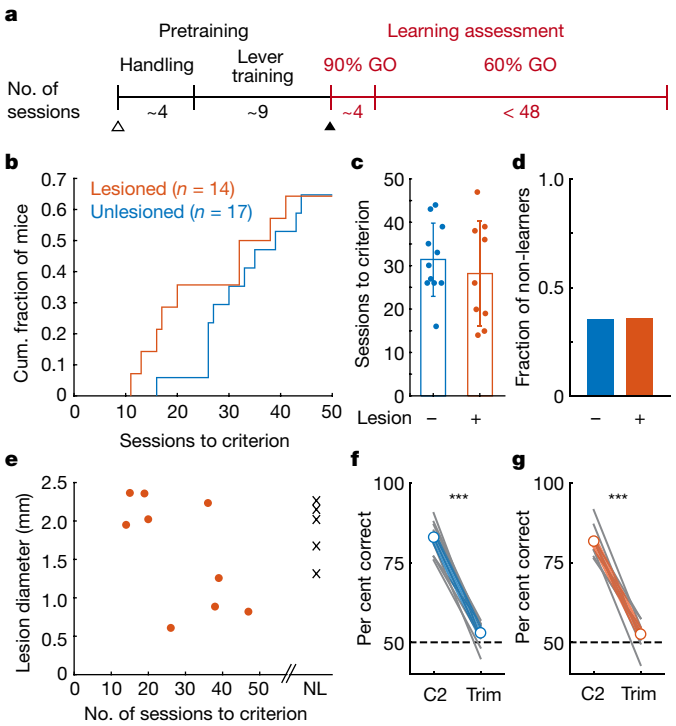


Fig. 4 | Barrel cortex is not required for learning whisker-dependent sensory detection task. a, Timeline for learning assay. **b**, Cumulative histogram of number of sessions to reach learning criterion (>74% correct for two consecutive sessions). The speed of learning was similar between unlesioned ($n = 17$) and lesioned ($n = 14$) mice. **c**, Among mice that learned the task, the number of sessions to criteria did not differ between unlesioned ($n = 11$) and lesioned ($n = 9$) mice. **d**, The fraction of mice that did not learn within the assessment period was similar between groups (unlesioned $n = 6$, lesioned $n = 5$ mice). **e**, Inability to learn was not due to larger lesion size. NL, non-learners. **f**, **g**, Unlesioned (**f**) and lesioned mice (**g**) that did learn could no longer perform the task when the C2 whisker was trimmed. ***$P < 0.001$.

Thus, three different manipulations of barrel cortex—Emx1–Halo, PV–ChR and lesions—transiently disrupt the execution of active detection behaviour. Recent studies have shown that the motor cortex might not be required for the execution of skilled movements, but is required for motor learning[27]. To determine whether the sensory cortex is required for learning of the detection task, we lesioned contralateral S1 of naive mice. Subjects were habituated and trained on lever manoeuvring. Learning rate was assessed, starting with the introduction to the pole (Fig. 4a, learning assessment, red). Initially, sessions consisted of 90% GO trials until mouse weight stabilized (3.6 ± 1.5 sessions), after which mice still performed at chance (Extended Data Fig. 7a–c). Lesioned and unlesioned mice learned at similar rates (Fig. 4b, c). Non-learners were equally present in both groups (Fig. 4d; 6/17 unlesioned, 5/14 lesioned, $P = 1$, Fisher's exact test), and failure to learn was not correlated with lesion size (Fig. 4e). In fact, among learners, mice with larger lesions learned faster than those with smaller lesions (Fig. 4e, linear regression, $P = 0.02$). Again, performance remained whisker-dependent (Fig. 4f, g). Thus, barrel cortex is not essential for learning the detection behaviour.

Task acquisition involves motor (lever press or lift), perceptual (pole detection) and contingency learning (lever→reward, contact→lever→reward). Notably, mice acquired the task whether they were lesioned before handling and lever training (Fig. 4a, open triangle, $n = 4$) or just before introduction of the pole (closed triangle, $n = 5$). Mice spent similar amounts of time in pre-training whether lesioned before pre-training or unlesioned ($P = 0.13$). Thus, barrel cortex appears not to be required for motor, perceptual and contingency learning of this task.

Our results demonstrate the potential for structures other than the sensory cortex to direct learned behaviours that require arbitrary coordination of multiple movements (lever press, whisking and licking) around a sensory event. This raises concerns about the interpretation of cortical physiology studies that use detection behaviours, as well as tasks requiring discrimination of elementary features encoded at the periphery (body location, retinotopic location, taste and sound frequency), which subjects may circumvent with detection strategies. It underscores the need to identify the behavioural conditions for which sensory cortex is indispensable, which might involve more complex discrimination, egocentric or allocentric context[10,28], or working memory[29].

In conclusion, impairment after transient inactivation does not absolutely indicate necessity. This raises the question of what the functional relevance of barrel cortex is to active detection. One possibility is that the barrel cortex and other structures are redundant for active detection[30]. Multiple subcortical structures receive barrel cortex input and, via other routes, whisker-related sensory signals. The trigeminal brainstem complex projects directly to the superior colliculus and cerebellum, and indirectly to the dorsolateral striatum via the secondary somatosensory thalamus[30–32]. Indeed, damage to striatum prevented recovery. Further studies are needed to assess the roles of other subcortical areas.

A second possibility is that manipulation of any cortical area may temporarily disrupt connected structures that are primarily involved in the task. In this scenario, the sudden loss of barrel cortex activity, rather than the sensory information it conveys, can be disruptive. The incomplete behavioural impairments we observed, as well as the sudden recovery after lesioning, raise the possibility of a disruptive effect, rather than redundancy. Subcortical systems are major targets of deep layer cortical pyramidal cells, which have high baseline firing rates, and removing their tonic activity may disrupt the responses of corticofugal targets to sensory inputs. In the birdsong motor system, lesions transiently disrupt activity in downstream areas and the production of song, both of which recover overnight[18]. In the birdsong study, it was unclear whether recovery was spontaneous or required some attempts, even if unsuccessful, at singing. A major advantage of our study is that we could control the time of re-exposure to the task after lesioning. Early task-specific experience accelerated recovery, and this may have important implications for early rehabilitation after stroke or head trauma. Whether recovery is always experience-dependent or whether sensory and motor systems differ are intriguing questions for further study.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0527-y.

Received: 28 September 2017; Accepted: 27 July 2018;
Published online 17 September 2018.

1. Glickfeld, L. L., Histed, M. H. & Maunsell, J. H. R. Mouse primary visual cortex is used to detect both orientation and contrast changes. *J. Neurosci.* **33**, 19416–19422 (2013).
2. Petruno, S. K., Clark, R. E. & Reinagel, P. Evidence that primary visual cortex is required for image, orientation, and motion discrimination by rats. *PLoS ONE* **8**, e56543 (2013).
3. Lashley, K. S. The mechanism of vision IV. The cerebral areas necessary for pattern vision in the rat. *J. Comp. Neurol.* **53**, 419–478 (1931).
4. Kato, H. K., Gillet, S. N. & Isaacson, J. S. Flexible sensory representations in auditory cortex driven by behavioral relevance. *Neuron* **88**, 1027–1039 (2015).
5. Kelly, J. B. & Glazier, S. J. Auditory cortex lesions and discrimination of spatial location by the rat. *Brain Res.* **145**, 315–321 (1978).
6. Talwar, S. K., Musial, P. G. & Gerstein, G. L. Role of mammalian auditory cortex in the perception of elementary sound properties. *J. Neurophysiol.* **85**, 2350–2358 (2001).
7. Oliveira-Maia, A. J. et al. The insular cortex controls food preferences independently of taste receptor signaling. *Front. Syst. Neurosci.* **6**, 5 (2012).
8. Peng, Y. et al. Sweet and bitter taste in the brain of awake behaving animals. *Nature* **527**, 512–515 (2015).
9. Waiblinger, C., Brugger, D. & Schwarz, C. Vibrotactile discrimination in the rat whisker system is based on neuronal coding of instantaneous kinematic cues. *Cereb. Cortex* **25**, 1093–1106 (2015).
10. Hutson, K. A. & Masterton, R. B. The sensory contribution of a single vibrissa's cortical barrel. *J. Neurophysiol.* **56**, 1196–1223 (1986).

© 2018 Springer Nature Limited. All rights reserved.

11. Morita, T., Kang, H., Wolfe, J., Jadhav, S. P. & Feldman, D. E. Psychometric curve and behavioral strategies for whisker-based texture discrimination in rats. *PLoS ONE* **6**, e20437 (2011).
12. Guo, Z. V. et al. Flow of cortical activity underlying a tactile decision in mice. *Neuron* **81**, 179–194 (2014).
13. O'Connor, D. H., Peron, S. P., Huber, D. & Svoboda, K. Neural activity in barrel cortex underlying vibrissa-based object localization in mice. *Neuron* **67**, 1048–1061 (2010).
14. Kwon, S. E., Yang, H., Minamisawa, G. & O'Connor, D. H. Sensory and decision-related activity propagate in a cortical feedback loop during touch perception. *Nat. Neurosci.* **19**, 1243–1249 (2016).
15. Miyashita, T. & Feldman, D. E. Behavioral detection of passive whisker stimuli requires somatosensory cortex. *Cereb. Cortex* **23**, 1655–1662 (2013).
16. Sachidhanandam, S., Sreenivasan, V., Kyriakatos, A., Kremer, Y. & Petersen, C. C. H. Membrane potential correlates of sensory perception in mouse barrel cortex. *Nat. Neurosci.* **16**, 1671–1677 (2013).
17. Stüttgen, M. C. & Schwarz, C. Barrel cortex: what is it good for? *Neuroscience* **368**, 3–16 (2018).
18. Otchy, T. M. et al. Acute off-target effects of neural circuit manipulations. *Nature* **528**, 358–363 (2015).
19. Deutsch, D., Pietr, M., Knutsen, P. M., Ahissar, E. & Schneidman, E. Fast feedback in active sensing: touch-induced changes to whisker-object interaction. *PLoS ONE* **7**, e44272 (2012).
20. Grant, R. A., Mitchinson, B., Fox, C. W. & Prescott, T. J. Active touch sensing in the rat: anticipatory and regulatory control of whisker movements during surface exploration. *J. Neurophysiol.* **101**, 862–874 (2009).
21. Mitchinson, B. et al. Active vibrissal sensing in rodents and marsupials. *Phil. Trans. R. Soc. Lond. B* **366**, 3037–3048 (2011).
22. Matyas, F. et al. Motor control by sensory cortex. *Science* **330**, 1240–1243 (2010).
23. Harvey, M. A., Sachdev, R. N. & Zeigler, H. P. Cortical barrel field ablation and unconditioned whisking kinematics. *Somatosens. Mot. Res.* **18**, 223–227 (2001).
24. Pammer, L. et al. The mechanical variables underlying object localization along the axis of the whisker. *J. Neurosci.* **33**, 6726–6741 (2013).
25. Stüttgen, M. C. & Schwarz, C. Psychophysical and neurometric detection performance under stimulus uncertainty. *Nat. Neurosci.* **11**, 1091–1099 (2008).
26. Keck, T. et al. Synaptic scaling and homeostatic plasticity in the mouse visual cortex *in vivo*. *Neuron* **80**, 327–334 (2013).
27. Kawai, R. et al. Motor cortex is required for learning but not for executing a motor skill. *Neuron* **86**, 800–812 (2015).
28. Brecht, M. The body model theory of somatosensory cortex. *Neuron* **94**, 985–992 (2017).
29. Stüttgen, M. C., Schwarz, C. & Jäkel, F. Mapping spikes to sensations. *Front. Neurosci.* **5**, 125 (2011).
30. Cohen, J. D. & Castro-Alamancos, M. A. Detection of low salience whisker stimuli requires synergy of tectal and thalamic sensory relays. *J. Neurosci.* **30**, 2245–2256 (2010).
31. Huerta, M. F., Frankfurter, A. & Harting, J. K. Studies of the principal sensory and spinal trigeminal nuclei of the rat: projections to the superior colliculus, inferior olive, and cerebellum. *J. Comp. Neurol.* **220**, 147–167 (1983).
32. Smith, J. B., Mowery, T. M. & Alloway, K. D. Thalamic POm projections to the dorsolateral striatum of rats: potential pathway for mediating stimulus-response associations for sensorimotor habits. *J. Neurophysiol.* **108**, 160–174 (2012).

Acknowledgements We thank B. C. Pil and A. Kase for assistance with mouse training, H. Zeng and M. Kheirbek for Ai39 mice, A. Losonczy for PV–Cre mice, J. Huang for Rosa–H2B–GFP mice, N. Clack for suggestions on video analysis of whisker motion, B. Grewe for advice on cortical aspiration, and J. C. Tapia, M. Goldberg, N. Sawtell, T. Jessell, A. Kinnischtzke, D. Kato and G. Pierce for comments on the manuscript. Funding was provided by NIH R01 NS094659, R01 NS069679, the Klingenstein Fund, the Rita Allen Foundation, the Dana Foundation and the Ludwig Schaefer Scholars Program (R.M.B.); NIH F32 NS084768 (Y.K.H.), T32 MH015174 (Y.K.H., C.O.L.) and F32 NS096819 (C.C.R.).

Reviewer information *Nature* thanks C. Schwarz and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.K.H. and R.M.B. conceived the experiments, analysed data and wrote the manuscript. Y.K.H. performed the experiments. C.O.L. and Y.K.H. developed the behavioural assay. C.C.R. assembled videography setup, C.C.R. and Y.K.H. performed array recordings, and C.C.R. analysed array data.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0527-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0527-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.M.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Transgenic mice. All mouse procedures complied with the NIH Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee at Columbia University. Emx1–Halo mice ($n = 10$) were generated by crossing Emx1-IRES-Cre[33] knock-in mice (Jackson Laboratories, stock #005628) to Rosa-lox-stop-lox (RSL)-eNpHR3.0/eYFP mice (stop-Halo Ai39, JAX, stock# 006364), which express halorhodopsin after excision of a stop cassette by Cre recombinase. Cre expression was assessed by crossing Emx1–Cre mice to RSL-H2B–GFP mice (provided by J. Huang). Negative control mice were Cre-negative and could not express the halorhodopsin transgene ($n = 7$ stop-Halo mice). PV–ChR mice ($n = 5$) were generated by crossing parvalbumin-Cre mice[34] to RSL-channelrhodopsin2/eYFP mice (Ai32, Jackson Laboratories, stock# 024109). For visualizing barrels in S1, Nr5a1–Cre mice (JAX, stock# 006364) were crossed to RSL-Halo–eYFP (Nr5a1–eYFP). All mouse lines were maintained on a C57BL/6 background. Optogenetic experiments used mice that were heterozygous for the desired transgene. The experimenters were not blind to genotype during testing and analysis.

Intrinsic signal optical imaging to locate C2 barrel. Mice were anaesthetized with isoflurane, and the skull over the left barrel cortex (centred ~1.5 mm posterior to bregma and 3.5 mm lateral to the midline) was thinned and sealed with Vetbond (3M) over a 4–5-mm area, or a glass window (3-mm coverslip, Warner Instruments) was implanted. Images were acquired with a CCD camera (Q-Imaging, Retiga 2000R) mounted on a stereomicroscope and software custom-written in LabVIEW. The vasculature on the brain surface was imaged with 510/40 band-pass filtered illumination (Chroma, D510/40), and functional imaging done with illumination with a 590-nm long-pass filter (Thorlabs, OG590). The C2 whisker was stimulated with 8 pulses of 4 directions at 5 Hz with a multi-directional piezo stimulator. The location of the maximum reflectance change was mapped relative to the surface vasculature. Two or three surrounding whiskers were also imaged to confirm proper identification of the C2 barrel location.

Behavioural setup. The behavioural setup was controlled by a microcontroller (Arduino), and data collected using custom-written routines. Subjects self-initiated trials by holding down a lever with their left forepaw for at least 100 ms. A pole (~2.15-mm-diameter wooden applicator stick) started from a position 3–4 cm below the mouse. After trial initiation, a stepper motor (Pololu Robotics and Electronics) rotated the pole to just in front of the whiskers (GO trials) or away from the whisker field entirely (NOGO trials). GO and NOGO trials were randomized. Rotation in either direction ensured that the sound and vibration generated by the motor was similar between trial types. The sound of the motor also served as a trial onset and offset cue. For GO trials, the pole was positioned 9–11 mm laterally and roughly aligned to the tip of the nose in the anterior–posterior dimension, such that mice were required to actively whisk forward to make contact. If the lever was lifted within the response window of 1.2 s during GO trials (hit), the response was rewarded with a drop of water and a reward tone. False alarm responses during NOGO trials were punished with a timeout period of 3–8 s accompanied by white noise sound, during which the mouse could not initiate a new trial. There was no reward for a correct reject and no punishment for a miss. The response latency was defined as the time from when the pole was first within reach of the mouse's whisker (typically 480 ms from trial onset, identified from high-speed video for each session analysed) to the mouse's lever lift response.

Mouse training. *Test of task performance.* Adult mice (P116 ± 60 days, mean and s.d., 34% male) were implanted with a custom-designed 22-gauge stainless steel laser-cut (Laser Alliance) headplate with dental acrylic. After ~1 week of recovery, subjects in optogenetics experiments were water-restricted and then trained in stages. Progression through each stage depended on the individual mouse's weight stabilization (indication of health). (1) Freely moving mice were habituated to the behavioural apparatus, where water reward was given for holding down a lever (2–3 days). (2) Mice were head-fixed and continued lever training: water was awarded for holding down lever for 500 ms to initiate trial, followed by releasing the lever for >100 ms (2–4 days). (3) Mice were trained in a dark chamber where no visual cues could be used. Mice were trained with 90% GO trials with a pole, and received water for responding by lifting a lever on GO trials, or a 3–5 s timeout if response was on a NOGO trial (2–8 days). We found that adding this gradual training stage with 90% GO trials facilitated learning. (4) Mice were trained at 60% GO trials until learning criterion (defined as >74% correct performance for 2 consecutive days) was reached.

Once subjects learned the detection task with all whiskers intact, the location of the C2 barrel was functionally mapped using intrinsic signal optical imaging. All whiskers except C2 on the mouse's right side were trimmed. Mice were retrained with a single C2 whisker until >74% performance was reached with 50% GO trials. In most cases, behavioural performance dropped after the initial whisker trimming but recovered over 1–7 days. Trimming was maintained twice a week. Lesions or sham operations were made after the performance of the mice stabilized above the performance criterion. Experimenters were not blind to whether mice were lesioned. Any animals that could still perform well (>60% correct) in the absence of all whiskers were excluded from analysis.

Test of task learning. Experimenters were blind to whether mice were lesioned or sham. For learning experiments, mice were trained as above but with more standardized training periods. Mice recovered from surgery and were habituated to handling by the trainer for 1–2 days. Mice were water-restricted and habituated through a series of pre-training stages. Progression through each step depended on the individual mouse's weight stabilization (indication of health), or whether rest days (weekends) interrupted training, rather than an explicit behavioural criterion. (1) Freely moving mice were habituated in the light to the behavioural apparatus, where water reward was given for holding down a lever (1–2 days). (2) Mice were head-fixed and continued to hold down the lever to drink (4 ± 2 days, mean and s.d.). (3) Mice were rewarded for holding down the lever and subsequently lifting for >100 ms (4 ± 2 days). All whiskers excluding C2 were trimmed twice a week for the remainder of the learning experiment. We quantified learning time for the following stages: mice were first introduced to the pole with 90% GO trials (~4 days, Extended Data Fig. 7). This stage and successive ones were performed in a light-tight box where no visual cues could be used. Mice were given <48 sessions to reach learning criterion of 74% correct for 2 consecutive sessions. To factor in the different number of GO and NOGO trials, we calculated per cent correct performance as $100 \times (N_{hit}/N_{GO} + N_{CR}/N_{nogo})/2$ in which the N variables are the numbers of hit, go, correct reject and NOGO trials, respectively.

A total of 29 mice were lesioned before learning. Of these, several mice were excluded from analysis, including 6 mice that performed >60% correct after full whisker trimming; 3 mice that were found to have lesions that did not include the C2 barrel; and 6 mice that had lesions that extended below the cortex, including white matter and striatum.

Quantification of whisker movements and contacts. Whisking was monitored with a high-speed camera (Photonfocus AG, MV1-D1312-100-G2) at 250 fps and 640 × 480 pixels/frame under infrared illumination. Whiskers were automatically traced offline and whisker position (angle) and curvature were obtained using Whisk[35,36]. Trials in which tracing failed >10% of frames were omitted from analysis.

For each session, mice were given a 5-min warm-up before analysis began, except for quantification of within-session performance (Fig. 3d) for which the entire session was included for analysis. Whisking analysis was restricted to a 200-ms window starting from the first frame in which whisker contact with the pole was possible. This window was chosen to best isolate the whisking associated with the sampling of the object and before the response, after which, whisking tended to increase in association with licking for water reward (Fig. 2g). For whisking amplitude and phase, the azimuthal whisking angle was band-pass filtered (four-pole Butterworth, 4–50 Hz) followed by a Hilbert transform[37,38]. Instantaneous frequency was calculated from the phase. The setpoint was measured as the midpoint between the whisking envelope defined by the maximum and minimum whisker angles for each whisk cycle.

Whisker contacts were defined using whisker angle and curvature parameters for each session. We first determined the range of angle-curvature values during free whisking in air (NOGO trials). The baseline curvature for each trial (mean change in curvature during 200-ms period before each trial onset) was subtracted to obtain the change in curvature (ΔK). Linear regression of the whisker angle and change in curvature was used to find the line of best fit (Extended Data Fig. 3). Upper and lower contact thresholds were set by finding the offset of the lines that encompassed the angle-curvature parameter space for all NOGO trials (1–5 s.d.). Putative contacts were defined as points at which local maxima or minima of ΔK were above or below the defined contact thresholds.

Optogenetic modulation of cortical activity. For optogenetic experiments during detection behaviour, the laser was on for 33–50% of trials, which were randomly interleaved with laser-off trials. For laser-on trials, an optical shutter opened at the onset of the trial, before movement of the pole. The pole moved within reach of the whisker field 200–400 ms after the onset of the laser, ensuring photoinactivation before contacts were possible. The laser remained on until after the mouse responded, or for the duration of the trial if no response was made, for a maximum of 1.5 s. Spiking activity during photoinhibition could be efficiently silenced for at least 2 s; all trials were 1.2–1.5 s in duration with a 1-s inter-trial-interval (Extended Data Fig. 1). Exposure to laser was limited to minimize photodamage to tissue. With the protocol described, no physiological or physical damage was detectable. A 593- or 594-nm laser (OEM or Coherent) was used with Emx1–Halo mice. For PV–ChR mice, an optical chopper (Thorlabs, MC2000B) modulated the 4 mW output of a 473-nm laser (OEM) to produce pulses at 40 Hz. Lasers were coupled to a 200-μm diameter, 0.39 NA optic fibre (Thorlabs) via a fibreport, and the diamond-knife cut fibre tip was placed above the optical window and positioned over C2 using the vasculature-referenced intrinsic signal map.

Electrophysiology. For recordings without behaviour, Emx1–Halo mice ($n = 4$) were habituated to head-fixation. Juxtasomal recordings were made using pipettes filled with artificial cerebral spinal fluid and an Axoclamp 900A amplifier. Airpuff

© 2018 Springer Nature Limited. All rights reserved.

stimuli for each test condition were delivered by opening an air valve for 50 ms during trials ranging from 0.5 to 2 s, and an inter-trial interval of 1.5–3 s for 30–50 trials. Laser-on trials were randomly interleaved for 50% of the trials. Glass pipettes were inserted perpendicular to the cortical surface (~30° from vertical), and the optic fibre was positioned vertically near the pipette entry point, above a thinned skull. To test the effect of photoinhibition at various distances, the optic fibre was positioned 0, 0.5, 1, 1.5 and 2 mm from the original recording site along a thinned and transparent skull ($n = 21$, 11, 10, 12 and 16, respectively). Regular spiking (RS, putative excitatory) versus fast-spiking (FS, putative inhibitory) cells were categorized based on their spike waveforms as previously described[39]. Cortical depth was defined as the microdrive depth relative to the pial surface.

For recordings during the behavioural task, a linear silicon array (Cambridge NeuroTech H3), consisting of 64 sites spanning 1,275 μm, was used. Each site was 11×15 μm and coated with PEDOT to obtain an impedance of 50–150 kΩ. Signals were band-pass filtered 1–7,500 Hz and sampled at 30 kHz (OpenEphys). Between sessions, the array was withdrawn and the craniotomy sealed with silicone. Spikes were clustered and inspected using Kilosort[40] and Phy[41].

Cortical lesions. Mice were deeply anaesthetized under isoflurane. A 1–4-mm craniotomy was made and the underlying cortical tissue was aspirated with a sterile blunt-tipped syringe needle connected to a vacuum. Lesions were made by aspirating all cortical layers to encompass, at a minimum, the C2 barrel and the immediately adjacent barrels, and at a maximum, the majority of S1 representing the large whiskers (macrovibrissae) and secondary somatosensory cortex (Extended Data Figs. 4, 5). Sham-operated mice were anaesthetized under the same conditions, and the skull was thinned with a dental drill. Lesioned and sham-operated mice were allowed to recover for 1 or 3 days after surgery before testing. After behavioural testing was complete, mice were perfused and brains extracted for histological analysis. Brains were sectioned tangentially or coronally (100 μm thick) with a vibratome. Lesion diameter was quantified in ImageJ by outlining the lesioned area for each section, quantifying the mean Feret diameter and averaging across all sections. Volume was measured by summing each section's lesion area multiplied by 100 (the section thickness).

In some cases, lesions extended beyond S1 and into subcortical tissue including the striatum. To objectively score striatal damage, the extent of cortical and subcortical damage was scored by five raters experienced at looking at coronal sections of mouse brains and blind to the behavioural data.

Data analysis. Analyses were done with custom-written scripts in MATLAB. For all figures, statistical significance is denoted as $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Differences were not significant ($P > 0.05$) unless otherwise indicated. Non-normally distributed data (D'Agostino–Pearson test) were reported using medians and interquartile ranges, and normally distributed data with means and s.e.m. Two-sided paired *t*-tests were used unless otherwise indicated. Based on the mean and s.d. of normal performance levels of trained mice, power analysis indicated that detecting a drop in performance to chance levels with a significance criterion of 0.05 required a minimum sample size of three, which we exceeded in all cases.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

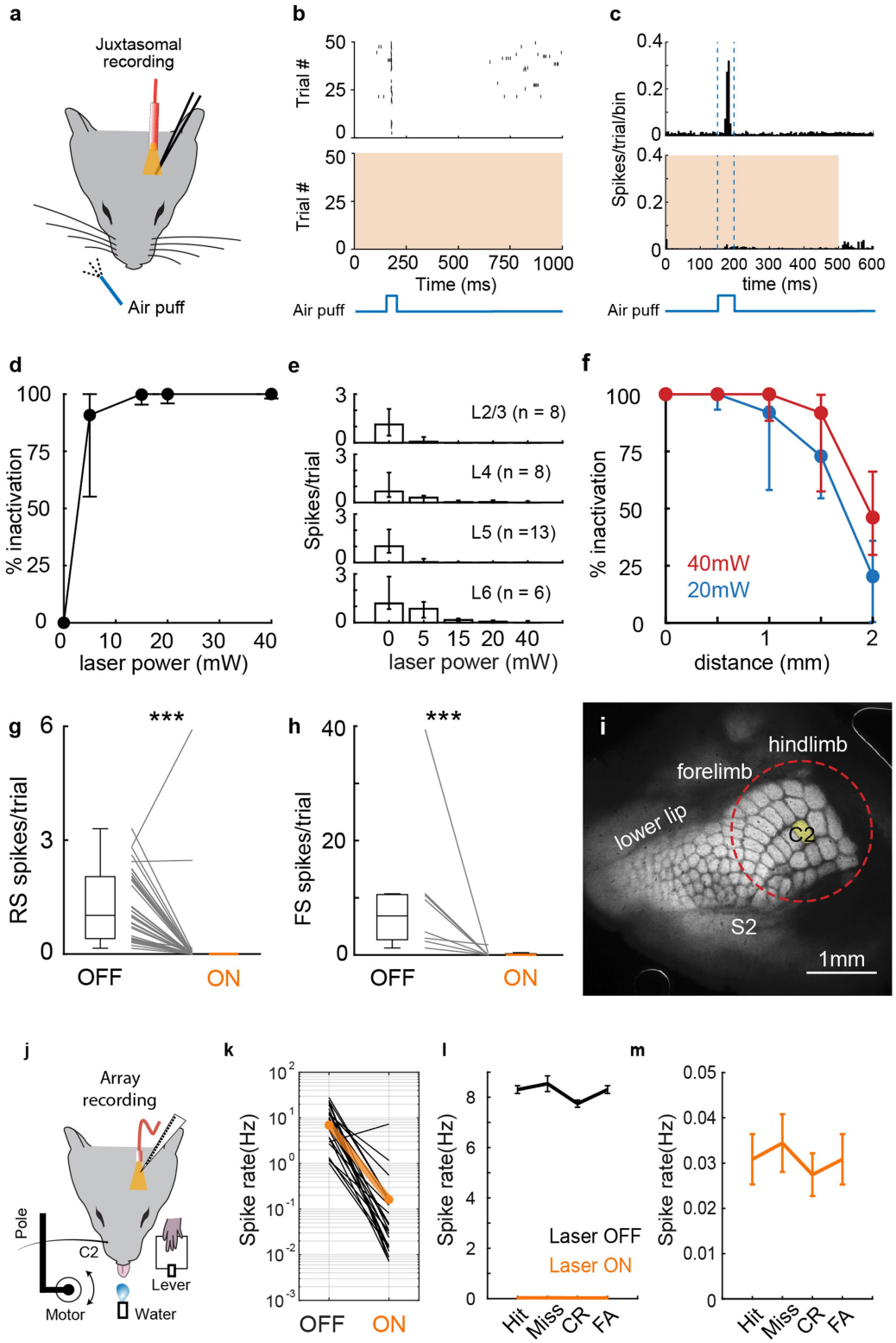
Code availability. All computer codes are available from the corresponding author upon reasonable request.

Data availability

All data are available from the corresponding author upon reasonable request.

33. Gorski, J. A. et al. Cortical excitatory neurons and glia, but not GABAergic neurons, are produced in the Emx1-expressing lineage. *J. Neurosci.* **22**, 6309–6314 (2002).
34. Scholl, B., Pattadkal, J. J., Dilly, G. A., Priebe, N. J. & Zemelman, B. V. Local integration accounts for weak selectivity of mouse neocortical parvalbumin interneurons. *Neuron* **87**, 424–436 (2015).
35. O'Connor, D. H. et al. Vibrissa-based object localization in head-fixed mice. *J. Neurosci.* **30**, 1947–1967 (2010).
36. Clack, N. G. et al. Automated tracking of whiskers in videos of head fixed rodents. *PLoS Comput. Biol.* **8**, e1002591 (2012).
37. Hill, D. N., Curtis, J. C., Moore, J. D. & Kleinfeld, D. Primary motor cortex reports efferent control of vibrissa motion on multiple timescales. *Neuron* **72**, 344–356 (2011).
38. Kleinfeld, D. & Deschênes, M. Neuronal basis for object location in the vibrissa scanning sensorimotor system. *Neuron* **72**, 455–468 (2011).
39. Bruno, R. M. & Simons, D. J. Feedforward mechanisms of excitatory and inhibitory cortical receptive fields. *J. Neurosci.* **22**, 10966–10975 (2002).
40. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Harris, K. D. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. Preprint at https://www.biorxiv.org/content/early/2016/06/30/061481 (2016).
41. Rossant, C. et al. Spike sorting for large, dense electrode arrays. *Nat. Neurosci.* **19**, 634–641 (2016).
42. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates*. 2nd ed. (Academic, New York, 2001).

© 2018 Springer Nature Limited. All rights reserved.

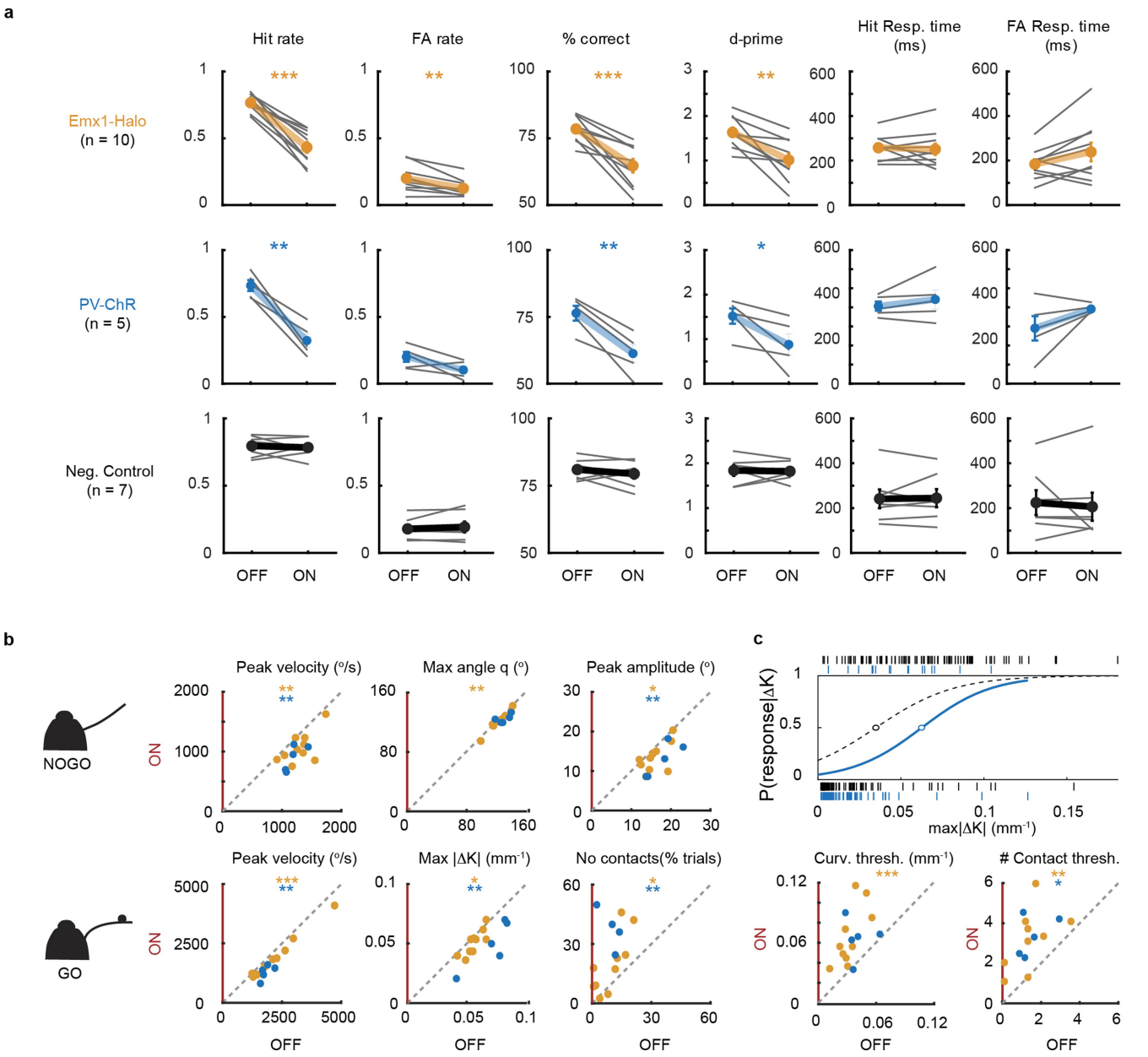


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

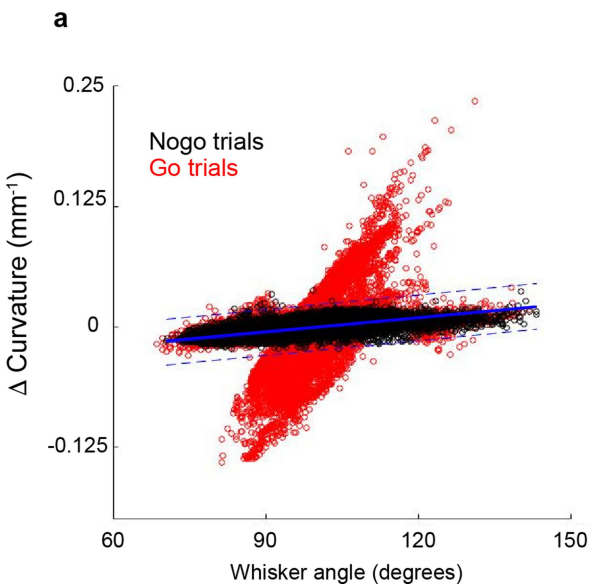
Extended Data Fig. 1 | Optogenetic photoinhibition of cortical neurons in Emx1–Halo mice is highly efficacious. a, Juxtasomal recordings in awake, head-fixed mice were made below an optic fibre placed above a thinned, transparent skull. **b**, Raster plot for example neuron for randomly interleaved laser off (top) or on (bottom) trials, with air puff schematized below. **c**, Population peri-stimulus time histograms of 35 cells with regular-spiking (RS) waveforms (cortical depth: 280–1,120 μm, $n = 4$ mice) for laser-off and -on trials. Both spontaneous and whisker stimulus-evoked spikes are silenced. Shaded area: laser on. **d**, **e**, Efficiency of RS cell inactivation as a function of laser power (**d**) and cortical depth (**e**), where per cent inactivation is relative to a cell's spike rate during laser-off trials. **f**, Lateral extent of inactivation. Illumination of 20–40 mW reliably inactivated an area within a 1-mm radius. **g**, Photoinhibition at 40 mW fully blocked spontaneous and sensory-evoked spikes (100% inactivation relative to laser-off trials) in 83% of RS cells and >96% of spikes in 94% of RS cells (same cells as in **c**); $P = 3.1 \times 10^{-11}$, Wilcoxon rank-sum test. **h**, Fast-spiking (FS) neurons ($n = 8$ cells) were similarly silenced; $P = 3.1 \times 10^{-8}$, Wilcoxon rank-sum test. **i**, Estimated area of photoinhibition with 40-mW relative to barrel cortex (1-mm radius around C2 barrel, red circle) depicted with a tangential section through barrel cortex of an Nr5a1–eYFP mouse with layer 4 labelled to visualize barrels. **j**, Emx1–Halo-mediated cortical inactivation was also assessed during detection behaviour with array recordings ($n = 8$ session, 3 mice). **k**, Data from Fig. 1e replotted on a logarithmic scale to show low spike rates during laser-on trials. **l**, Behavioural performance during laser-off and -on trials did not correlate with spiking activity for each trial type (four sessions from three mice with per cent correct for laser-off trials: 82, 87, 80, and 78%). **m**, Laser-on data in **l**, replotted with larger scale to visualize data during laser-on trials. Data shown as median ± interquartile range (**d**–**h**); mean ± s.e.m. (**l**, **m**).

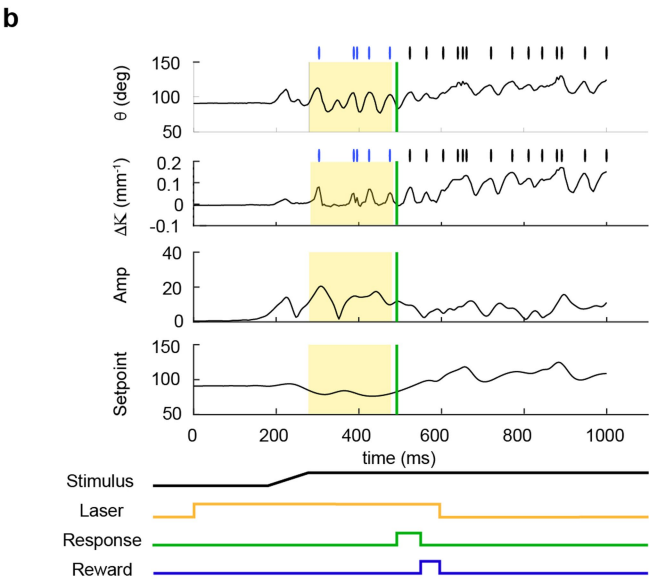
© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | Optogenetic manipulations of barrel cortex using PV–ChR and Emx1–Halo mice result in similar behavioural, motor, and sensory deficits. a, Photoinhibition of excitatory neurons in Emx1–Halo mice (orange) and photoactivation of inhibitory neurons in PV–ChR mice (blue) yielded similar behavioural deficits. Negative control mice (Cre-negative, stop-Halo mice; black) were unaffected by 593-nm laser illumination. *P* values for Emx1–Halo: hit, 1.6×10^{-5}; false alarm (FA), 1.7×10^{-3}; per cent correct, 3.6×10^{-4}; d-prime, 1.8×10^{-3}. For PV–ChR: hit, 3.5×10^{-3}; per cent correct, 3.9×10^{-3}; d-prime, 0.0498. **b**, Optogenetic inactivation of S1 with either Emx1–Halo or PV–ChR mice decreases whisking kinematics. *P* values for Emx1–Halo: NOGO peak velocity, 6.6×10^{-3}; max angle, 1.4×10^{-3}; peak amplitude, 2.6×10^{-2}; GO peak velocity, 4.3×10^{-4}; max change in curvature (ΔK), 1.9×10^{-2}; per cent trials with no contacts, 1.2×10^{-2}. PV–ChR: NOGO peak velocity, 8.1×10^{-3}; max angle, 8.7×10^{-2}; peak amplitude, 8.4×10^{-2}; GO peak velocity, 5.7×10^{-3}; ΔK, 8.7×10^{-3}; per cent trials with no contacts, 7.5×10^{-3}. **c**, Sensory thresholds increase with optogenetic inactivation. *P* values for Emx1–Halo: curvature threshold, 9.7×10^{-4}; contact threshold, 3.4×10^{-3}. PV–ChR: curvature threshold, 1.1×10^{-1}; contact threshold, 1.2×10^{-2}. Data for negative control and Emx1–Halo mice are the same as in Figs. 2, 3 but repeated here for comparison with PV–ChR mice.

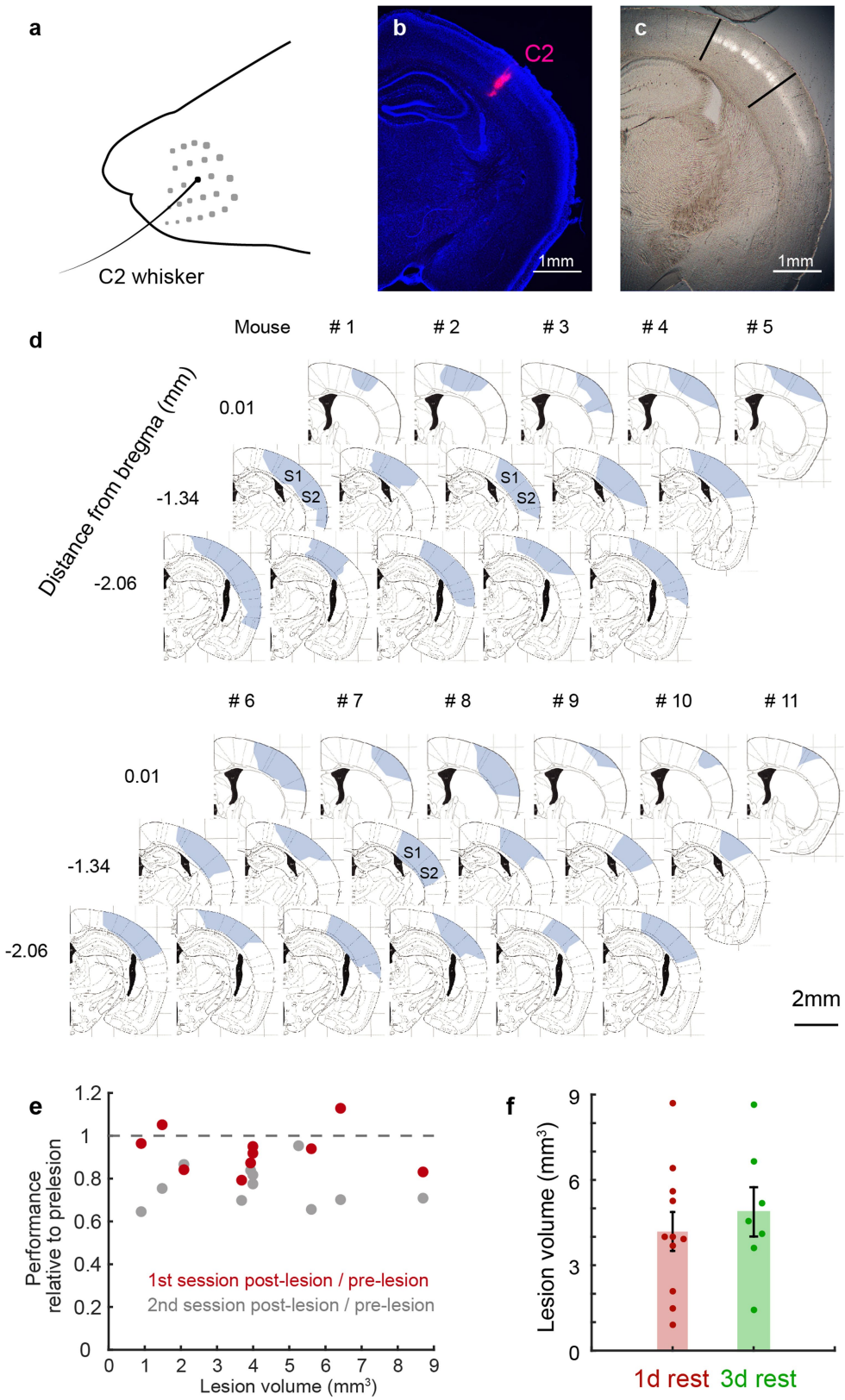
© 2018 Springer Nature Limited. All rights reserved.





Extended Data Fig. 3 | Defining contacts based on whisker angle and change in curvature. a, Example curvature versus whisker position for a single session. Each circle represents the paired values for curvature and whisker angle for each frame during the session. Values for NOGO trials define whisker parameters during free whisking in air, when no contacts can be made (black); GO trials are shown in red. Linear regression was used to define the line of best fit (blue, solid line) for NOGO parameters, and upper and lower contact thresholds were set by finding the offsets that encompassed the no-contact parameter space (1–5 s.d. from the line of best fit, blue dashed lines). **b**, Putative contacts were defined as points at which the local maxima or minima of the change in curvature were above (forward contact with whisker) or below (reverse contact with whisker) the defined thresholds (tick marks). Whisking analysis was restricted to the 200-ms time window (yellow shaded area) during sampling, before the average response time (green).

© 2018 Springer Nature Limited. All rights reserved.

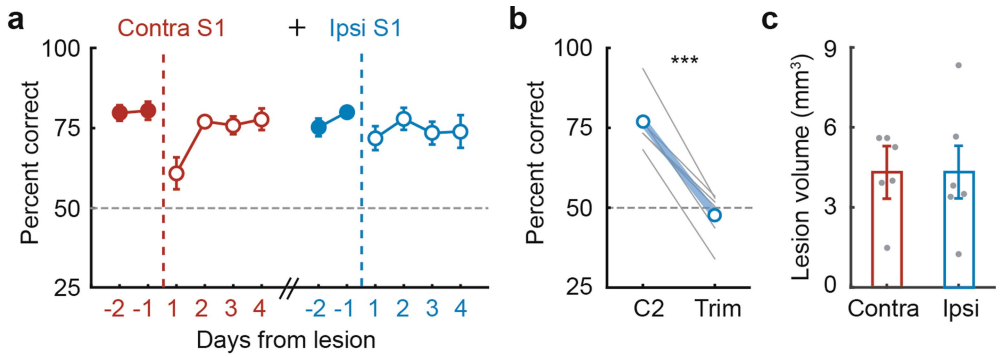


Extended Data Fig. 4 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

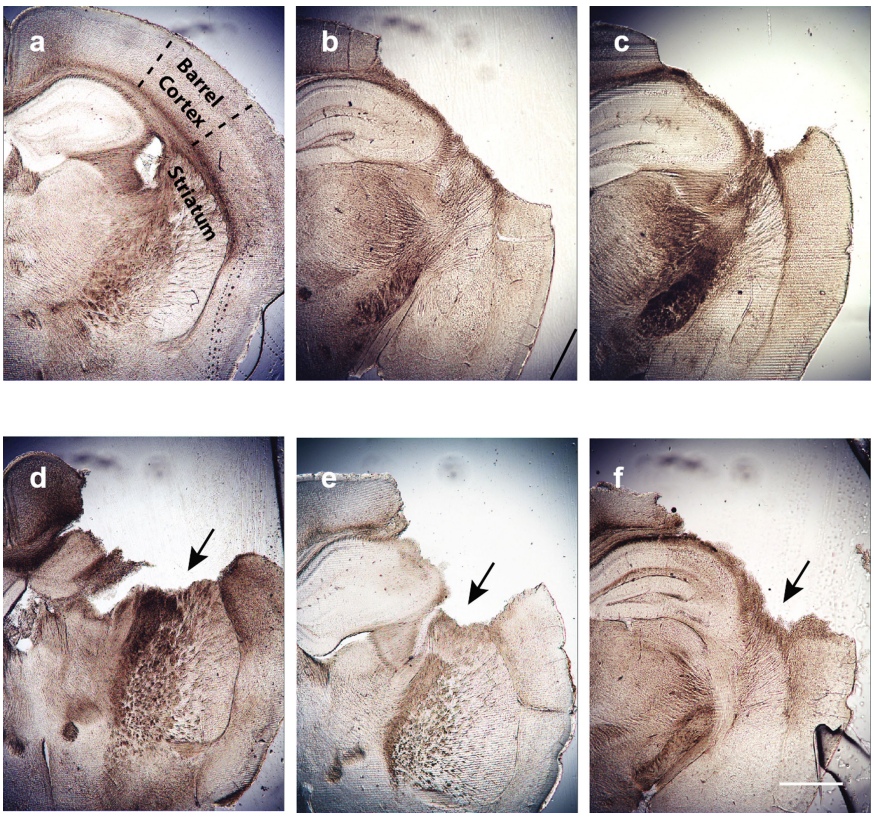
Extended Data Fig. 4 | Behavioural performance after lesions did not correlate with lesion size. a, Mice performed the task with a single C2 whisker. **b**, The location of the C2 barrel in a coronal section. The C2 barrel was functionally mapped with intrinsic imaging and Alexa-conjugated cholera toxin subunit B (CTB, red) was injected into the centre of the C2 barrel. Blue, DAPI. Mappings in **b** and **c** are used to inform the locations of lesions made relative to the C2 barrel column. **c**, Equivalent location in section from an Nr5a1–eYFP mouse with barrels fluorescently labelled in L4 (white) overlaid on bright-field image to show extent of barrel cortex relative to section (black lines). C2 was located about 1.2–1.5 mm posterior to bregma, varying slightly between mice. Lesions were centred around C2. **d**, Size and locations of contralateral barrel cortex lesions for the 11 mice with 1-day rest shown in Fig. 3 (arranged from largest to smallest by lesion volume). For each mouse, three locations along the anterior–posterior axis are shown overlaid on atlas images, reproduced with permission from ref. [42]. In a few mice (for example, mice 1, 3 and 8), lesions extended into the secondary somatosensory area (S2). Numbers along anterior–posterior axis indicate approximate location relative to bregma. **e**, Lesion size did not correlate with the degree of impairment on the first (grey) or second post-lesion session when behavioural performance had recovered (red). Performance was normalized to the pre-lesion performance for each mouse. **f**, Lesion sizes were similar between groups with 1 or 3 days of rest after lesioning ($P = 0.91$).

© 2018 Springer Nature Limited. All rights reserved.



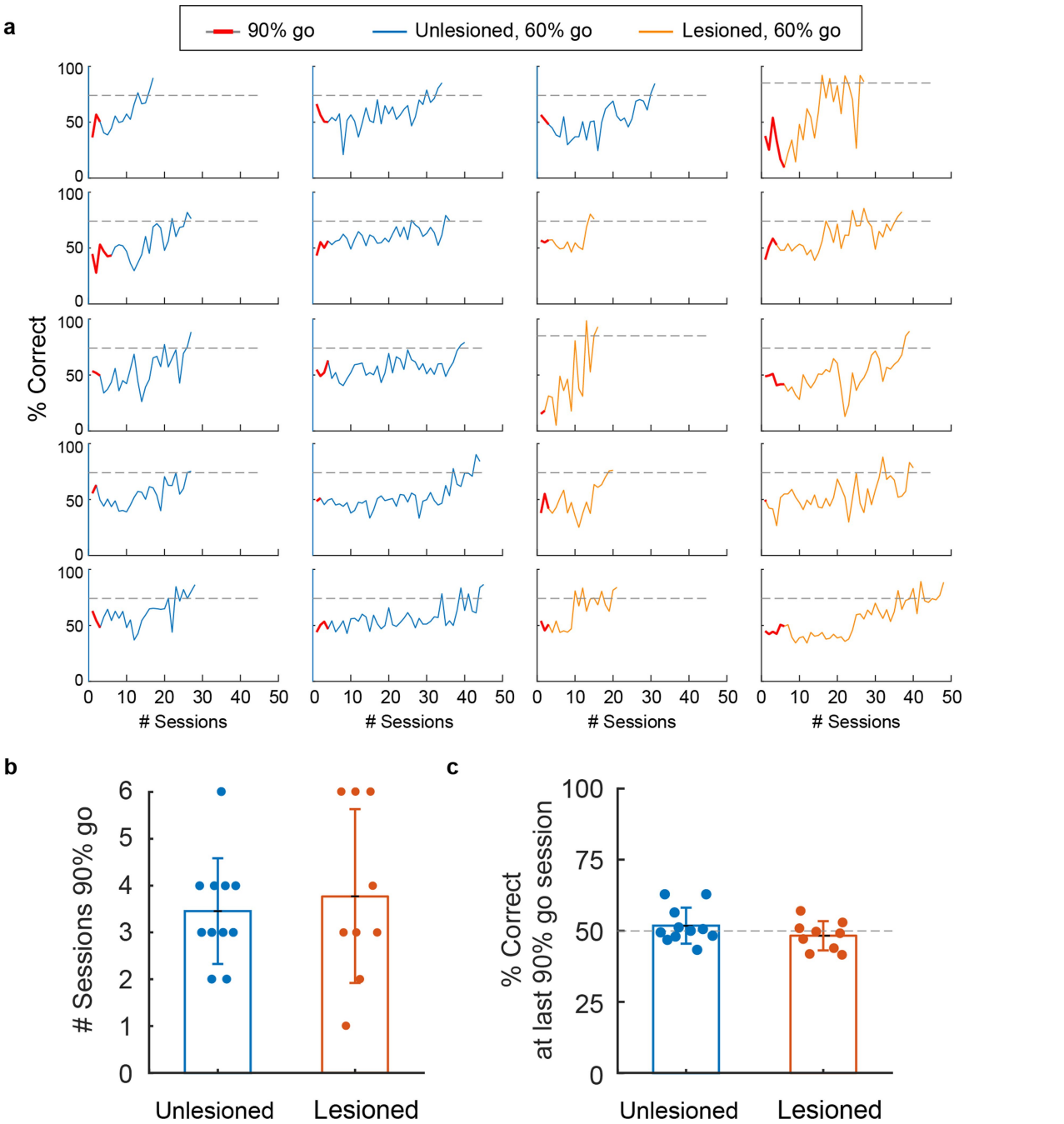
Extended Data Fig. 5 | Ipsilateral S1 does not compensate for loss of contralateral S1. a, Behavioural performance of mice recovered rapidly after contralateral S1 lesions, as shown in Fig. 3c (red). Subsequent ipsilateral lesion ($n = 6$) had effects similar to sham and ipsilateral-only manipulations shown in Fig. 3c, indicating that ipsilateral S1 was not compensating for loss of contralateral S1 activity. **b**, C2 whisker-trim control. Performance of bilaterally lesioned mice dropped to chance when the C2 whisker was removed ($P = 2.7 \times 10^{-3}$). **c**, Sizes of bilateral lesions.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Example histology from lesioned mice depicting lesion of contralateral S1 only or additional damage to striatum. **a**, Unlesioned example. **b**, **c**, Examples of S1-only lesions from two mice. **d**–**f**, Three examples of damage to striatum in addition to S1. Even minimal damage (arrows) to the dorsolateral striatum resulted in permanent behaviour deficits. Scale bar, 1 mm.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Learning curves for unlesioned and lesioned mice in learning experiment. a, Individual learning curves for unlesioned mice ($n = 11$, blue lines) and mice with lesions of contralateral barrel cortex ($n = 9$, orange lines) that learned the detection task to criterion (74% correct performance for two consecutive sessions). Mice were first introduced to the pole with 90% GO trials (red lines). This intermediate step ensured that mice maintained a stable weight before moving on to the last step of training. Mice were moved onto 60% GO trials for the rest of the learning assessment. Mice were given 48 sessions to learn the task. **b**, Unlesioned and lesioned groups spent similar times on 90% GO sessions (3.8 ± 1.8 versus 3.5 ± 1.1 sessions for unlesioned and lesioned mice, respectively; $P = 0.64$). **c**, By the end of the 90% GO sessions, performance was still at chance levels ($P = 0.34$ and $P = 0.37$ for unlesioned and lesioned mice, respectively; one-sample t-test).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0515-2

Touch and tactile neuropathic pain sensitivity are set by corticospinal projections

Yuanyuan Liu[1,2,9], Alban Latremoliere[1,3,9], Xinjian Li[4,5,9], Zicong Zhang[1,2,9], Mengying Chen[1,6], Xuhua Wang[1], Chao Fang[1], Junjie Zhu[1], Chloe Alexandre[3,7], Zhongyang Gao[1], Bo Chen[1,2], Xin Ding[1], Jin-Yong Zhou[1], Yiming Zhang[1], Chinfei Chen[1,2], Kuan Hong Wang[4,10]*, Clifford J. Woolf[1,2,8,10]* & Zhigang He[1,2,10]*

Current models of somatosensory perception emphasize transmission from primary sensory neurons to the spinal cord and on to the brain[1–4]. Mental influence on perception is largely assumed to occur locally within the brain. Here we investigate whether sensory inflow through the spinal cord undergoes direct top-down control by the cortex. Although the corticospinal tract (CST) is traditionally viewed as a primary motor pathway[5], a subset of corticospinal neurons (CSNs) originating in the primary and secondary somatosensory cortex directly innervate the spinal dorsal horn via CST axons. Either reduction in somatosensory CSN activity or transection of the CST in mice selectively impairs behavioural responses to light touch without altering responses to noxious stimuli. Moreover, such CSN manipulation greatly attenuates tactile allodynia in a model of peripheral neuropathic pain. Tactile stimulation activates somatosensory CSNs, and their corticospinal projections facilitate light-touch-evoked activity of cholecystokinin interneurons in the deep dorsal horn. This touch-driven feed-forward spinal–cortical–spinal sensitization loop is important for the recruitment of spinal nociceptive neurons under tactile allodynia. These results reveal direct cortical modulation of normal and pathological tactile sensory processing in the spinal cord and open up opportunities for new treatments for neuropathic pain.

To investigate whether descending corticospinal projections regulate sensory processing in the spinal cord, we performed a bilateral pyramidotomy[6,7] to lesion the CST in the brainstem of adult mice (Fig. 1a, b) and tested them for sensitivity to noxious and innocuous sensory stimuli (Fig. 1c–i), with histological verification (Extended Data Fig. 1a). Bilateral CST lesions caused a marked deficit in behavioural responses to dynamic and static light touch, as indicated by a reduced response to stroking with a soft paintbrush (Fig. 1h) and increased latency to detect tape on the hindpaw plantar surface (Fig. 1i). Similarly, in the von Frey test, mice with CST lesions exhibited reduced responses to innocuous low pressure punctate stimuli[8] (0.16 g) but had normal sensitivity to higher mechanical forces (0.6–2 g), which activate nociceptors[9] (Fig. 1g). The hindpaw withdrawal responses to noxious thermal and mechanical stimuli were intact in mice with a pyramidotomy (Fig. 1c–f), as was gross motor performance (Extended Data Fig. 1b, c). The selective deficits in light touch sensitivity (Extended Data Fig. 1a) corroborate the finding that CST axons co-terminate with low-threshold mechanosensory afferents in laminae III–V[10,11] (Extended Data Fig. 2a, b).

To determine which areas of the cerebral cortex[12] contain CSNs that may be responsible for the sensory modulation, we used anterograde tracing to identify innervation of the dorsal horn of the spinal cord. CST axons from hindlimb primary sensory (S1) cerebral cortex, with a minor contribution from secondary sensory cortex (S2), terminate in laminae III–V (Fig. 2a, b), suggesting that they have a sensory role, whereas the hindlimb motor cortex projects more ventrally, to laminae VI–VII (Fig. 2a, b), consistent with its known motor functions. Thus, CST axons from S1 and S2 are more likely to execute control over tactile sensory processing in the spinal cord. To test this further, we used intersectional targeting[12–14] to express human diphtheria toxin receptor (DTR) in S1/S2 CSNs, enabling selective ablation of these neurons.

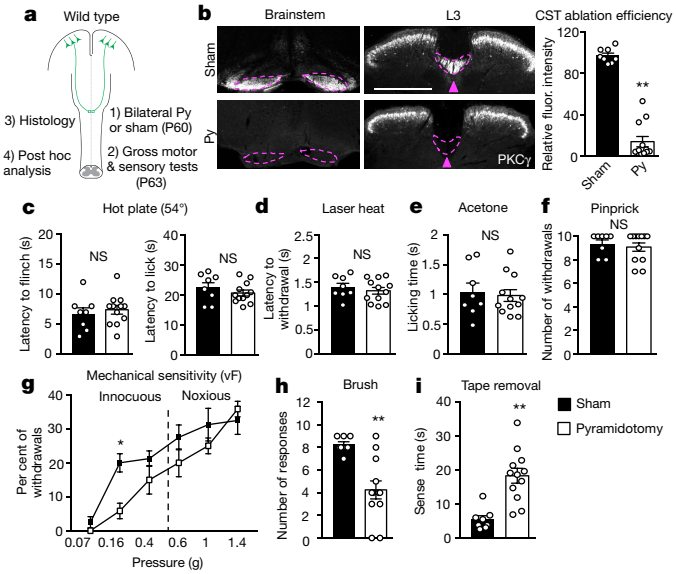


Fig. 1 | Adult CST ablation impairs light touch but not nociceptive behavioural responses. a, Schematic of experimental paradigm. Py, pyramidotomy. **b**, Left, representative images of transverse sections of the brain stem at pyramids and dorsal spinal cord (L3) in sham ($n = 8$) or pyramidotomized ($n = 12$) mice used for sensory tests stained with anti-PKC$_\gamma$. Arrowheads and dashed lines indicate the location of the main CST. Scale bar, 500 μm. Right, quantification of PKC$_\gamma$ immunofluorescence intensity (normalized to control) in the lumbar dorsal funiculus of sham or pyramidotomized mice. $P < 0.0001$. **c**–**i**, Measurement of thermal and mechanical sensitivities by hot plate (**c**, $P = 0.59$ and 0.35 for flinch and lick, respectively), laser heat (**d**, $P = 0.50$), acetone (**e**, $P = 0.73$), pinprick (**f**, $P = 0.58$), von Frey filaments (vF; **g**, $P = 0.016$ for 0.16 g), dynamic light brush (**h**, $P = 0.001$), and tape removal (**i**, $P = 0.0003$) in sham ($n = 8$) or pyramidotomized mice ($n = 12$). NS, not significant. **b**, **c**–**f**, **h**, **i**, Two-sided *t*-test; **g**, two-way repeated measures ANOVA followed by Bonferroni correction. Data shown as mean ± s.e.m.

[1]F. M. Kirby Neurobiology Center and Department of Neurology, Boston Children's Hospital, Boston, MA, USA. [2]Department of Neurology, Harvard Medical School, Boston, MA, USA. [3]Neurosurgery Department, Johns Hopkins School of Medicine, Baltimore, MD, USA. [4]Unit on Neural Circuits and Adaptive Behaviors, Clinical and Translational Neuroscience Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA. [5]Department of Neurology and Research Center of Neurology in Second Affiliated Hospital, Key Laboratory of Medical Neurobiology of Zhejiang Province, Zhejiang University School of Medicine, Hangzhou, China. [6]Department of Child Health Care, Children's Hospital of Nanjing Medical University, Nanjing, China. [7]Beth Israel Deaconess Medical Center, Boston, MA, USA. [8]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [9]These authors contributed equally: Yuanyuan Liu, Alban Latremoliere, Xinjian Li, Zicong Zhang. [10]These authors jointly supervised this work: Kuan Hong Wang, Clifford J. Woolf, Zhigang He. *e-mail: wkuan@mail.nih.gov; clifford.woolf@childrens.harvard.edu; zhigang.he@childrens.harvard.edu

© 2018 Springer Nature Limited. All rights reserved.

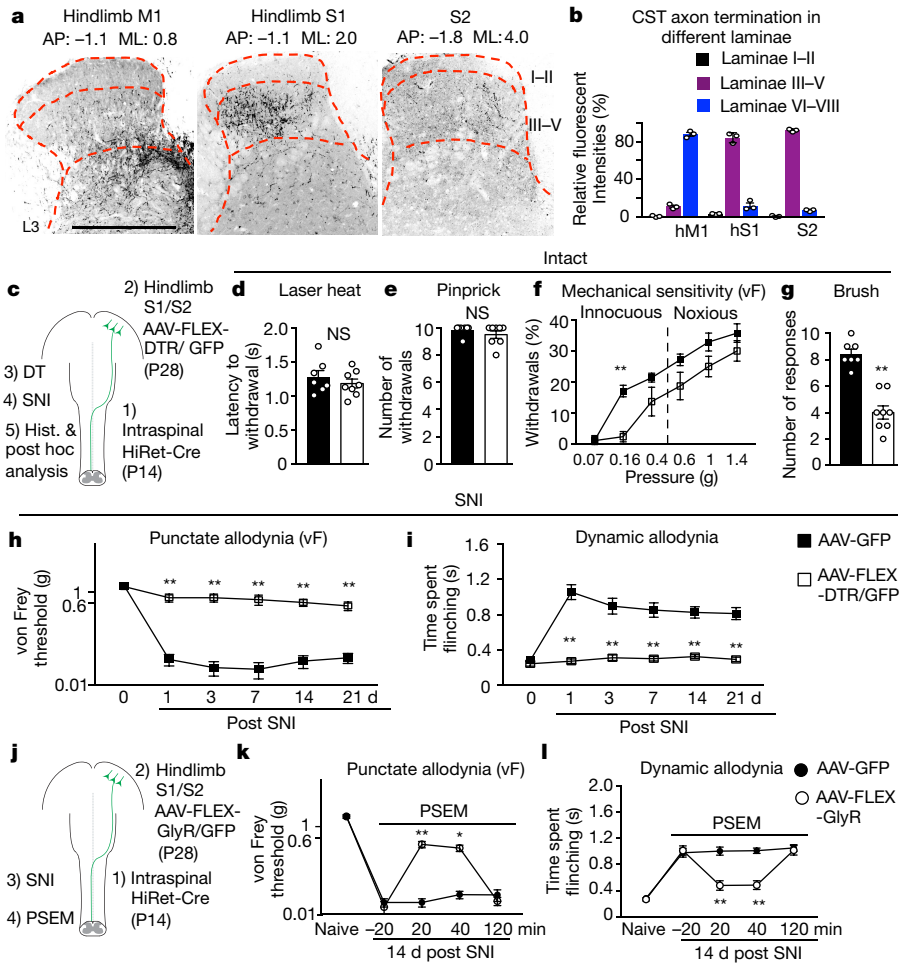


Fig. 2 | Hindlimb somatosensory CSN ablation causes specific loss of light touch response and mechanical allodynia. a, Representative images showing projection patterns of CST axons originating from indicated cortical areas. AP, anteroposterior; ML, mediolateral. Scale bar, 500 μm. **b**, Quantification of relative CST termination percentile within different laminae from indicated cortical areas ($n = 3$). **c**, Schematic of regional CSN ablation where HiRet viruses were injected into the lumbar spinal cord (T13–L6) at P14. DT, diphtheria toxin; SNI, spared nerve injury. **d**–**g**, Measurements of sensitivity to laser heat (**d**, $P = 0.40$), pinprick (**e**, $P = 0.28$), von Frey filaments (**f**, $P = 0.004$ for 0.16g) and brush stimuli (**g**, $P < 0.0001$) in control mice ($n = 7$) and mice in which hindlimb somatosensory CSNs have been ablated ($n = 8$). **h**, **i**, Measurement of punctate and dynamic mechanical allodynia after SNI in control mice ($n = 7$) and mice in which hindlimb somatosensory CSNs have been ablated ($n = 8$). **j**, Schematic of silencing of S1/S2 CSNs by injection of HiRet-Cre into the lumbar spinal cord at P14. PSEM, pharmacologically selective effector molecule (a specific ligand to activate engineered GlyR expressed in the hindlimb S1/S2). **k**, **l**, Measurement of punctate and dynamic allodynia after SNI with PSEM administration in mice injected with AAV-FLEX-GlyR ($n = 7$) or AAV-GFP ($n = 8$). $P = 0.0001$ and 0.0006 for 14 and 21 d, respectively, in **i**; $P = 0.002$ and 0.016 for 20 and 40 min, respectively, in **k**; elsewhere, $^{**}P < 0.0001$. **d**, **e**, **g**, Two-sided *t*-tests; **f**, **h**, **i**, **k**, **l**, two-way repeated measures ANOVA followed by Bonferroni correction. Data shown as mean ± s.e.m.

High-efficiency retrograde lentiviral vectors (HiRet) expressing Cre were injected into the lumbar spinal cord (thoracic (T)13–lumbar (L)6) of mice at postnatal day (P)14, a time when cortical pruning is complete[15,16], followed by an injection of AAV-FLEX-DTR into S1/S2 (Fig. 2c). Administration of diphtheria toxin to these mice ablated about 85% of CSNs in S1/S2 (Extended Data Fig. 3), and mice with such selective ablation fully phenocopied the results of pyramidotomy (Fig. 2d–g), indicating that the CST axons from S1/S2 cortex contribute to tactile sensitivity.

A prominent feature of peripheral neuropathic pain is mechanical allodynia, in which light touch triggers pain[3,4]. To assess whether the CST modulates mechanical allodynia, we subjected mice with ablation of S1/S2 CSNs to the spared nerve injury (SNI) model of neuropathic pain[17]. Mice in which S1/S2 CSNs had been ablated exhibited a profound reduction in both the punctate allodynia evoked by low-intensity static von Frey stimuli and in dynamic allodynia evoked by brushes at all time points tested (1 to 21 days after SNI) (Fig. 2h, i). Mice with a pyramidotomy also failed to develop mechanical allodynia induced by SNI (Extended Data Fig. 4a–c) or by intraplantar injection of complete Freund's adjuvant (CFA), a model of peripheral inflammatory pain (Extended Data Fig. 4f, g). However, the CST-lesioned mice still exhibited cold allodynia and hyperalgesia in response to intense mechanical stimuli after SNI (Extended Data Fig. 4d, e). We conclude that CSNs from S1/S2 are required for the induction of mechanical allodynia but not other aspects of the pain hypersensitivity phenotype. In support of this, transiently silencing CSNs in the hindlimb S1/S2 cortex using an inducible chemogenetic method[18] diminished established mechanical allodynia in mice with SNI (Fig. 2j–l). These results show that somatosensory CSNs projecting to the deep dorsal horn contribute to both innocuous tactile processing in normal conditions and the generation of mechanical allodynia in pathological states, something not suspected before.

To assess how CSNs in S1 respond to peripheral tactile stimulation, we used in vivo miniaturized microscopy[12,19] to monitor calcium levels in CSNs expressing GCaMP6s, a genetically encoded fluorescent calcium indicator (Fig. 3a, Extended Data Fig. 5). S1 CSNs in intact freely moving mice showed sparse and temporally scattered activity (Fig. 3b, c). By contrast, brush or von Frey stimulation, but not laser heat, to the hindpaw elicited calcium activity in a subset of S1 CSNs, and such activation was significantly increased after SNI in the same mice

© 2018 Springer Nature Limited. All rights reserved.

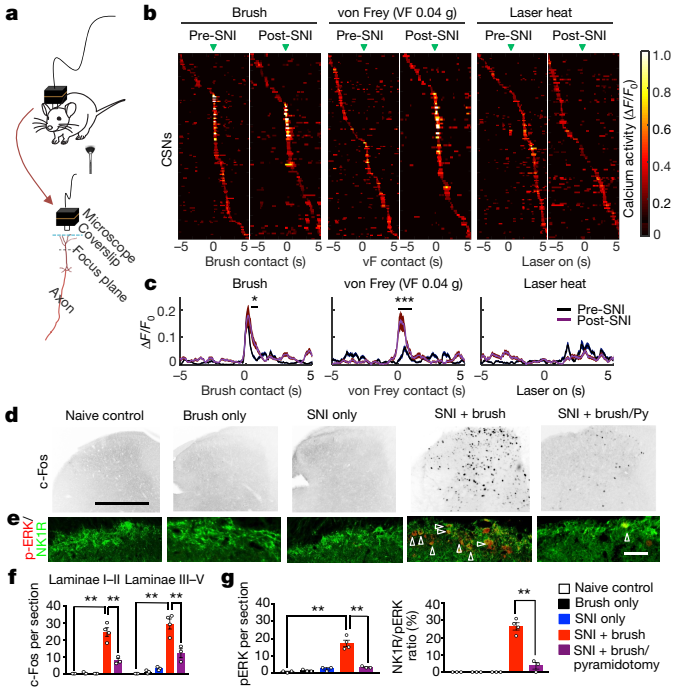


Fig. 3 | Light-touch-elicited activation of somatosensory CSNs and spinal dorsal horn neurons. a, Diagram of calcium activity imaging in S1 CSNs with a head-mounted miniaturized microscope. **b**, Trial-averaged activity of CSNs in response to indicated sensory stimuli before and after SNI. Active calcium event traces are aligned to the time when the brush (pre-SNI, $n = 156$ neurons from 4 mice; post-SNI, $n = 103$ neurons from 4 mice), von Frey filament (pre-SNI, $n = 151$ neurons from 4 mice; post-SNI, $n = 94$ neurons from 4 mice), or laser (pre-SNI, $n = 156$ neurons from 4 mice; post-SNI, $n = 116$ neurons from 4 mice) affected the hindpaw (green arrowheads). Trial averages (13–17 trials per mouse) were sorted by peak activity time. Arrowheads indicate onset of stimulation. **c**, Average response to brush, von Frey filament (0.04 g) and laser heat stimuli before and after SNI. Number of neurons as in **b**. $*P = 0.03$ and $***P = 0.0002$ for brush and von Frey, respectively; rank sum test. **d**–**g**, Representative images (**d**, **e**) and quantification (**f**, **g**) of c-Fos and pERK (red) or NK1R (green) immunoreactivity in the dorsal horn of the spinal cord (L3–4) in mice under the following conditions: naive control ($n = 3$), brush only ($n = 3$), SNI only ($n = 3$), SNI and brush ($n = 4$), SNI and brush with pyramidotomy ($n = 3$). Scale bars, 500 and 100 μm in **d**, **e**, respectively. Arrowheads indicate co-localization of pERK and NK1R. **f**, **g**, One-way ANOVA with Bonferroni correction. $P = 0.0008$ for SNI + brush with or without pyramidotomy in **f**; otherwise, $**P < 0.0001$. Data shown as mean ± s.e.m.

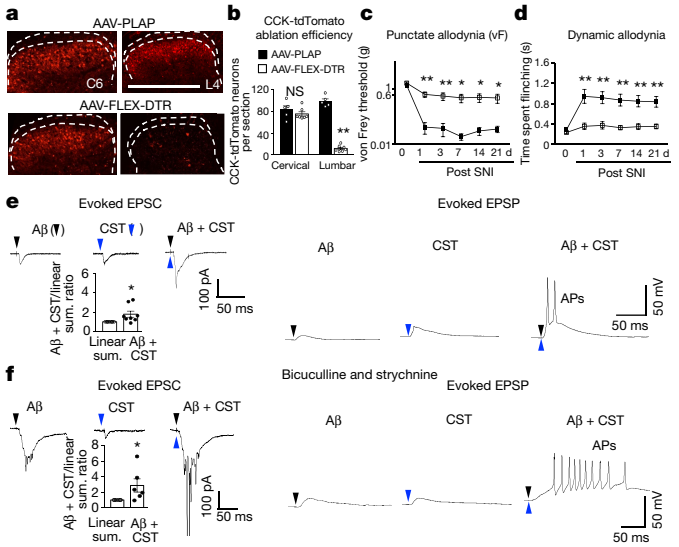


Fig. 4 | Lumbar CCK$^+$ neurons receive convergent Aβ fibre and CST inputs and are required for mechanical allodynia. a, **b**, Representative images (**a**) of cervical and lumbar spinal cord dorsal horn in control mice (AAV-PLAP, $n = 5$) or mice in which CCK–tdTomato neurons are ablated (AAV-FLEX-DTR, $n = 7$) with quantification (**b**). $P = 0.35$ and $P < 0.0001$ for cervical and lumbar, respectively; two-sided *t*-test. **c**, **d**, Measurement of punctate (**c**) and dynamic (**d**) mechanical allodynia after SNI in control mice ($n = 8$) or mice in which CCK–tdTomato interneurons have been ablated ($n = 7$). Two-way repeated measures ANOVA followed by Bonferroni correction. $P = 0.002$, 0.01, 0.03, 0.03, 0.04 and $P < 0.0001$, $P = 0.0003$, 0.0003, 0.0009, 0.0014 for 1, 3, 7, 14, 21 d in **c** and **d**, respectively. **e**, **f**, Averaged traces of EPSCs evoked by Aβ, CST, or co-stimulation (−70 mV, voltage clamp, left) and representative trace of EPSP or action potential (AP; current clamp, right) in untreated CCK–tdTomato neurons (**e**), or neurons treated with bicuculline and strychnine (**f**). Bar graphs under EPSC curves: ratio of EPSC amplitude evoked by Aβ and CST co-stimulation over the sum of amplitudes evoked by individual stimulation (**e**, $n = 8$, $P = 0.02$; **f**, $n = 6$, $P = 0.04$), one-sided paired *t*-test. Data shown as mean ± s.e.m.

(Fig. 3b, c). In addition, the proportion of cells that responded to both brush and von Frey stimulation was increased after SNI (Extended Data Fig. 5). Thus, CSNs in the somatosensory cortex of intact mice are activated by innocuous tactile stimuli and this is amplified under neuropathic pain conditions, suggesting a spinal–cortical–spinal sensitization loop.

We next examined whether descending CST axons modulate the neuronal activation evoked by low-intensity mechanical stimulation in the spinal cord. We detected few c-Fos$^+$ cells in the dorsal horn of intact mice after brush stimulation or in mice with SNI without brush stimulation (Fig. 3d). However, brush stimulation in mice with SNI increased the number of c-Fos$^+$ cells in all laminae (I–V) of the ipsilateral dorsal spinal cord (Fig. 3d, f). A CST lesion reduced this effect in both deep laminae (III–V) and superficial laminae (I/II), a region that normally processes only inputs from nociceptors and which is not contacted directly by CST (Fig. 3d, f). These results suggest that CST inputs contribute to the level of brush-evoked neuronal activation in all laminae of the dorsal horn. Neurons expressing the neurokinin 1 receptor NK1R in lamina I of the spinal cord are a major nociceptive output pathway to the brain[20,21]. To assess whether CSNs modulate the patterns of activation of these neurons, we monitored phospho-ERK (pERK) immunoreactivity as an activity marker[22]. Tactile stimulation in mice with SNI, but not in intact mice, induced a significant, CST-dependent increase in the number of pERK–NK1R$^+$ neurons in lamina I (Fig. 3e, g). Tactile stimulation after SNI in the presence of an intact CST directly activates neurons in laminae III–V and indirectly increases activation of lamina I projection neurons. The latter effect is likely to result from disinhibition of polysynaptic circuits that connect laminae III–V with lamina I[4,20,23]. CST ablation blocked the mechanical allodynia caused by spinal disinhibition produced pharmacologically by inhibiting GABA (γ-aminobutyric acid) and glycine activity[21,23] (Extended Data Fig. 6). These data indicate that inputs from the corticospinal pathway to their direct targets in laminae III–V are required for low-threshold afferent inputs to activate ascending nociceptive pathways after SNI. Consistent with this conclusion, the activation of multiple brain areas, including CSNs in S1/S2, produced by tactile stimulation after SNI is blocked by CST ablation (Extended Data Fig. 7).

We next investigated which neurons in the dorsal horn are targeted by the S1/S2 CST. Excitatory cholecystokinin (CCK)^{iresCre}-labelled cells in the deep dorsal horn are co-innervated by the CST and Aβ primary afferent fibres, and selective silencing of CCK$^+$ neurons results in tactile defects[11], making it likely that CCK neurons are CST targets. Indeed, selective ablation of CCK$^+$ interneurons in the lumbar spinal cord (T13–L6) phenocopied CST lesions (Figs. 1, 2), producing selective defects in tactile sensitivity in intact mice and a reduction in mechanical allodynia after SNI (Fig. 4a–d, Extended Data Fig. 8a–f). Ablation of CCK$^+$ interneurons also significantly reduced the number of c-Fos$^+$ cells after brush stimulation in all dorsal horn laminae in mice with SNI (two-sided *t*-test, $P = 0.0045$ and 0.0028 for laminae I–II and

© 2018 Springer Nature Limited. All rights reserved.

laminae III–V, respectively, Extended Data Fig. 8g, i), but to a lesser extent than pyramidotomy (Fig. 3d, f), suggesting that other neurons co-innervated by CST and Aβ fibres also contribute to this process[11]. Nevertheless, these results show that CCK^+ interneurons in deep laminae of the dorsal horn are important contributors to the facilitatory effects of the CST on tactile sensitivity and mechanical allodynia.

To test whether CCK^+ interneurons are directly activated by the CST and Aβ fibres, we performed whole-cell patch recordings in ex vivo spinal cord slices[21,24,25] with attached dorsal roots prepared from CCK–tdTomato mice[26] in which S1/S2-derived CST axons were labelled with channelrhodopsin2 and yellow fluorescent protein (ChR2–YFP) (Extended Data Fig. 9a). Among all recorded CCK–tdTomato cells, 50% exhibited excitatory monosynaptic responses to both Aβ and CST stimulation (Extended Data Fig. 9b, c), but in only a few was single stimulation of either input sufficient to evoke action potentials (Extended Data Fig. 9b, c). However, in those neurons with direct connections from Aβ fibres and the CST (Extended Data Fig. 9d), co-stimulation of the afferents and descending pathway led to super-additive synaptic responses, with a substantially increased action potential output (Fig. 4e). In the presence of bicuculline (a $GABA_A$ receptor inhibitor) and strychnine (a glycine receptor inhibitor), which mimics the spinal disinhibition that occurs in SNI[21,23], such additive responses were further enhanced, with action potential trains occurring in all recorded CCK–tdTomato neurons (Fig. 4f). In mice with SNI, half of $c\text{-}Fos^+$ cells found in the dorsal horn after repetitive touch stimulation expressed CCK (Extended Data Fig. 8g), and c-Fos immunoreactivity in these cells was abolished by pyramidotomy (Extended Data Fig. 8g, h). Together, these results suggest that CST inputs onto CCK^+ interneurons have a crucial sensitizing role in the processing of tactile input in naive mice, and that this role is increased after nerve injury.

Finally, we found that optical stimulation of CSNs in S1 in mice expressing ChR2 targeted to CSNs (Extended Data Fig. 10a) significantly extended the duration of paw flinching or licking after a light brush stimulus (two-sided *t*-test, $P = 0.02$, Extended Data Fig. 10b, c), and increased brush-induced pERK in lamina I NK1R neurons (Extended Data Fig. 10d). We conclude that the direct output from the somatosensory cortex to the spinal dorsal horn via the CST mediates powerful facilitation of tactile sensory processing in the spinal cord, which amplifies transmission of tactile-related signalling to the brain. This represents a spino-cortico-spinal feed-forward sensitization loop that is crucial for controlling tactile sensation in normal conditions and allodynia in neuropathic pain states. Normalizing the excitability of somatosensory CSNs could, therefore, be a potential target for treating neuropathic pain by pharmacological or electromagnetic manipulation. Furthermore, by identifying direct cortical control of tactile processing within spinal circuits, these results provide a plausible explanation for how mental states could directly increase or decrease normal and pathological tactile sensations in different contexts or mood states, by controlling activity in transmission pathways from the spinal cord to the brain.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0515-2.

Received: 17 October 2017; Accepted: 23 July 2018;
Published online 12 September 2018.

1. Owens, D. M. & Lumpkin, E. A. Diversification and specialization of touch receptors in skin. *Cold Spring Harb. Perspect. Med.* **4**, a013656 (2014).
2. Abraira, V. E. & Ginty, D. D. The sensory neurons of touch. *Neuron* **79**, 618–639 (2013).
3. Basbaum, A. I., Bautista, D. M., Scherrer, G. & Julius, D. Cellular and molecular mechanisms of pain. *Cell* **139**, 267–284 (2009).
4. Costigan, M., Scholz, J. & Woolf, C. J. Neuropathic pain: a maladaptive response of the nervous system to damage. *Annu. Rev. Neurosci.* **32**, 1–32 (2009).
5. Lemon, R. N. Descending pathways in motor control. *Annu. Rev. Neurosci.* **31**, 195–218 (2008).
6. Liu, K. et al. PTEN deletion enhances the regenerative ability of adult corticospinal neurons. *Nat. Neurosci.* **13**, 1075–1081 (2010).
7. Jin, D. et al. Restoration of skilled locomotion by sprouting corticospinal axons induced by co-deletion of PTEN and SOCS3. *Nat. Commun.* **6**, 8074 (2015).
8. Boada, M. D. & Woodbury, C. J. Myelinated skin sensory neurons project extensively throughout adult mouse substantia gelatinosa. *J. Neurosci.* **28**, 2006–2014 (2008).
9. Boada, M. D. & Woodbury, C. J. Physiological properties of mouse skin sensory neurons recorded intracellularly *in vivo*: temperature effects on somal membrane properties. *J. Neurophysiol.* **98**, 668–680 (2007).
10. Lemon, R. N. & Griffiths, J. Comparing the function of the corticospinal system in different species: organizational differences for motor specialization? *Muscle Nerve* **32**, 261–279 (2005).
11. Abraira, V. E. et al. The cellular and synaptic architecture of the mechanosensory dorsal horn. *Cell* **168**, 295–310.e19 (2017).
12. Wang, X. et al. Deconstruction of corticospinal circuits for goal-directed motor skills. *Cell* **171**, 440–455.e14 (2017).
13. Kinoshita, M. et al. Genetic dissection of the circuit for hand dexterity in primates. *Nature* **487**, 235–238 (2012).
14. Liu, Y. et al. A sensitized IGF1 treatment restores corticospinal axon-dependent functions. *Neuron* **95**, 817–833.e4 (2017).
15. O'Leary, D. D. Development of connectional diversity and specificity in the mammalian brain by the pruning of collateral projections. *Curr. Opin. Neurobiol.* **2**, 70–77 (1992).
16. O'Leary, D. D. & Koester, S. E. Development of projection neuron types, axon pathways, and patterned connections of the mammalian cortex. *Neuron* **10**, 991–1006 (1993).
17. Decosterd, I. & Woolf, C. J. Spared nerve injury: an animal model of persistent peripheral neuropathic pain. *Pain* **87**, 149–158 (2000).
18. Atasoy, D., Betley, J. N., Su, H. H. & Sternson, S. M. Deconstruction of a neural circuit for hunger. *Nature* **488**, 172–177 (2012).
19. Li, X. et al. Skin suturing and cortical surface viral infusion improves imaging of neuronal ensemble activity with head-mounted miniature microscopes. *J. Neurosci. Methods* **291**, 238–248 (2017).
20. Todd, A. J. Neuronal circuitry for pain processing in the dorsal horn. *Nat. Rev. Neurosci.* **11**, 823–836 (2010).
21. Cheng, L. et al. Identification of spinal circuits involved in touch-evoked dynamic mechanical pain. *Nat. Neurosci.* **20**, 804–814 (2017).
22. Ji, R. R., Baba, H., Brenner, G. J. & Woolf, C. J. Nociceptive-specific activation of ERK in spinal neurons contributes to pain hypersensitivity. *Nat. Neurosci.* **2**, 1114–1119 (1999).
23. Torsney, C. & MacDermott, A. B. Disinhibition opens the gate to pathological pain signaling in superficial neurokinin 1 receptor-expressing neurons in rat spinal cord. *J. Neurosci.* **26**, 1833–1843 (2006).
24. Duan, B. et al. Identification of spinal circuits transmitting and gating mechanical pain. *Cell* **159**, 1417–1432 (2014).
25. Bourane, S. et al. Identification of a spinal circuit for light touch and fine motor control. *Cell* **160**, 503–515 (2015).
26. Taniguchi, H. et al. A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron* **71**, 995–1013 (2011).

Acknowledgements We thank T. Huang, Y. Zhang and Q. Ma for advice and D. Ginty, S. Hegarty, Q. Ma, F. Wang and P. Williams for critical reading. This study was supported by grants from the Craig Neilsen Foundation (Y.L. and X.W.), Paralyzed Veterans of America Foundation (Y.L.), Dr. Miriam and Sheldon G. Adelson Medical Research Foundation and NINDS (C.J.W. and Z.H.) and NIMH intramural research program ZIA MH002897 (K.H.W. and X.L.). IDDRC and viral cores supported by the grants NIH P30 HD018655 and P30EY012196 were used for this study.

Reviewer information *Nature* thanks R. Ji and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.L., A.L., X.L., Z.Z., K.H.W., C.J.W. and Z.H. conceived the experiments. Y.L., A.L., X.L., Z.Z., M.C., X.W., C.F., C.A., J.Z., Z.G., B.C., X.D., J.-Y.Z. and Y.Z. performed the experiments. Y.L., A.L., X.L., C.C., K.H.W., C.J.W. and Z.H. prepared the manuscript with input from all authors. K.H.W., C.J.W. and Z.H. co-supervised the project.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0515-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0515-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.H.W. or C.J.W. or Z.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Mouse strains. All experimental procedures were performed in compliance with animal protocols approved by the Institutional Animal Care and Use Committee at Boston Children's Hospital or National Institute of Mental Health. C57Bl/6 wild-type, Rosa-LSL-ChR2EYFP (Jax#12569), Rosa-LSL-tdTomato[27] and CCK^{Cre} (Jax#012706) mouse strains were maintained on a C57Bl/6 genetic background. For each set of behavioural measurements, animals used in experimental and control groups were littermates. The body weight and sexes were randomized and assigned to different treatment groups, and no other specific randomization was used for the animal studies. Behavioural tests were run blinded to the experimental condition. No statistical methods were used to predetermine sample size.

Virus production and intraspinal injection. Lentivirus vectors of HiRet-GFP and HiRet-Cre (all lenti-virus titres were adjusted to 1.6–2 $\times 10^{12}$ copies/ml for injection) were constructed based on the HiRet backbone[13]. To retrogradely target supraspinal neurons, HiRet lentiviruses were injected into the mouse spinal cord guided by a Vevo 770 ultrasound system (VisualSonics) as described[28].

Regional ablation or silencing of corticospinal neurons or CCK$^+$ interneurons. For regional CSN ablation, 2 μl HiRet-Cre viruses (2×10^{12} copies/ml) were intraspinally injected into wild-type mice at P4 (pilot experiment, data not shown) or P14. 1 μl AAV2/8-FLEX-DTR viruses (5×10^{12} copies/ml) or AAV2/8-FLEX-PLAP viruses (5×10^{12} copies/ml) as control were injected to the cortical areas covering ACC/PFC (AP: 0.8–2.5 mm, ML: 0.2–0.8 mm to bregma) or the somatosensory cortex (S1: −0.3 to −1.6 mm, ML: 1.5–2.4 mm to bregma; S2: AP: −1.8 to −2.1 mm, ML: 3.3–4.0 mm to bregma), respectively. To monitor injection position and targeting efficiency, all mice were co-injected with AAV2/8-GFP into the cortex (5×10^{12} copies/ml). Diphtheria toxin (DT, Sigma, D0564, 100 μg/kg) was administered intraperitoneally into adult animals (age 8 weeks). After behavioural and histological assessments were performed blindly, mice with incomplete CST ablation, as defined by more than 25% spared GFP$^+$ axons in the lumbar spinal cord, were excluded from further analysis. In our pilot experiment (P4 injection), we found that DT-mediated CSN ablation in S1/S2, but not PFC/ACC, significantly impaired responses to light touch, whereas the responses to noxious heat or mechanical stimuli were not affected by these manipulations.

To transiently silence CSN activity, 2 μl HiRet-Cre or PLAP viruses (2×10^{12} copies/ml) were injected to the lumbar spinal cord at P14 or 1 μl AAV-2/8-FLEX-PSAM$^{L141F\text{-}Y115F}$-GlyR-IRES-GFP viruses (5×10^{12} copies/ml) were injected into the cortical areas covering hindlimb S1/S2 cortical areas at P28. After von Frey and brush measurements (see 'Behavioural assessment') after spared nerve ligation (SNI), PSEM^{308} (Apex Scientific, 10 mg/kg) was administered intraperitoneally. von Frey and brush measurements were performed at 20, 40 and 120 min post PSEM administration. We verified the expression (GFP$^+$) of GlyR in the somatosensory cortex as post hoc histological analysis.

To ablate lumbar spinal CCK$^+$ interneurons, we injected 3 μl AAV2/8-FLEX-DTR viruses (5×10^{12} copies/ml) or AAV2/8-FLEX-PLAP viruses (5×10^{12} copies/ml) as control at P21 into the lumbar spinal cord (T13–L6) of CCK–tdTomato mice. DT (Sigma, D0564, 100 μg/kg) was administered intraperitoneally at P35. Animals were subjected to gross motor and sensory tests 2 weeks after DT injection. For post hoc examination, CCK–tdTomato$^+$ cells were quantified in the cervical and lumbar spinal cord in mice injected with AAV-FLEX-DTR or AAV-FLEX-PLAP, respectively.

Immunohistochemistry and imaging. Mice were perfused transcardially with 4% paraformaldehyde (PFA) in phosphate-buffered saline (PBS) followed by further 4% PFA in PBS post-fixation overnight. The tissue was cryo-protected with 30% sucrose and processed using a cryostat (section thickness 30 μm for spinal cord and 60 μm for brain). Sections were intensively washed with PBS and then treated with a blocking solution containing 10% goat normal serum (GNS), 0.5% Triton-100 for 2 h at room temperature before staining. The primary antibodies used were chicken anti-GFP [Abcam (ab13970), 1:1,000], rabbit anti-RFP [Abcam (ab34771), 1:1,000], rabbit anti-PKCγ [Santa Cruz (sc211), 1:100], rabbit anti-c-fos [Cell Signaling Technology (2250), 1:200], mouse anti-pERK [Cell Signaling Technology (9106), 1:100], and rabbit-anti-NK1R [Sigma-Aldrich (S8305), 1:1,000]. Secondary antibodies including Alexa Fluor 488-conjugated goat anti-chicken, Alexa Fluor 488, 594-conjugated goat anti-rabbit, and the fluorescent IB4 (Alex Fluor 488 conjugated) were purchased from Invitrogen. Floating brain tissue sections (60 μm) were incubated with primary antibodies overnight at 4 °C and washed three times for 10 min with PBS before being incubated with secondary antibodies, or with IB4 (1:400 dilution) for 2 h. After thorough washing with PBS, sections were mounted with ProLong antifade mounting medium with or without DAPI for imaging and analysis. Sections were imaged with a confocal laser-scanning microscope (Zeiss 700), or a wide field Nikon 80i Microscope (air).

Surgery. For pyramidotomy, the procedure is similar to that described previously[8,9]. In brief, mice were anaesthetized with ketamine and xylazine. An incision was made at the left side of the trachea, followed by blunt dissection to expose the skull base. A craniotomy in the occipital bone was then performed to allow access to the medullary pyramids. The left and right pyramids were cut with a fine scalpel medially up to the basilar artery. The wound was closed in layers with 6.0 sutures. After pyramidotomy, mice were tested for gross motor functions. Most animals showed normal gross motor functions, consistent with previous findings[9,12,29,30]. At the end point, lumbar spinal cord sections from all mice were subjected to immunostaining with antibodies against PKCγ and the remaining PKCγ immunofluorescence in the dorsal column was quantified. As shown in Extended Data Fig. 1, a few mice with inefficient CST ablation showed neither nociceptive nor tactile sensation deficits. With these criteria, we excluded animals with inefficient CST ablation and/or gross motor defects from all other independent experiments (CFA, c-fos, pERK quantification).

For anterograde labelling of CST axons (Fig. 2a, b), AAV2/1-ChR2-mCherry (1×10^{13} copies/ml for injection, produced by Boston Children's Hospital, viral core) was injected into mouse sensorimotor cortex at multiple sites as described previously[12].

Spared nerve ligation (SNI) surgery was performed under isoflurane (3% for induction /2% for maintenance) on adult mice (8–12 weeks old). In brief, the tibial and common peroneal branches of the sciatic nerve were tightly ligated with a 5.0 silk suture and transected distally, while the sural nerve was left intact[15]. After injury, the incision was sutured and mice were allowed to recover on heated pads before being returned to their home cages. Punctate and dynamic mechanical allodynia, cold allodynia and mechanical hyperalgesia were measured 1–21 days after SNI. To induce spinal disinhibition, we intrathecally injected 10 μl saline containing bicuculline (0.02 μg, Sigma-Aldrich, 14340) and strychnine (0.05 μg Sigma-Aldrich, S0532) into the spinal cord[21]. We started to measure mechanical allodynia 10 min after injection. For CFA-induced inflammation, mice were anaesthetized with isoflurane (2%). 20 μl of CFA (Sigma) was then injected into the plantar surface of the left hindpaw. Both punctate and dynamic mechanical allodynia were measured 1–7 days after CFA treatment.

Behavioural experiments. *Contact heat pain (hot plate test).* Mice were placed on a metallic plate heated to a set temperature (30, 50 or 54 °C) within an acrylic container (Bioseb), and videotaped. The latency to flinching and licking one of the hindpaws was blindly measured by video analysis. One temperature was tested per day.

Laser heat pain. Individual mice were habituated for 20 min in a small (7.5 × 7.5 × 15 cm) plastic cage. A laser heat source (500 mW, 445 nm; Shanghai Laser & Optics Century) was targeted at the paw and the latency to withdraw was measured.

Acetone test (cold allodynia). Mice were habituated for 20 min in a small (7.5 × 7.5 × 15 cm) plastic cage and a small volume of acetone (5 μl) was applied onto the plantar surface of the hindpaw. The time spent flinching or licking the paw was recorded for one minute.

Pinprick. Mice were habituated for 20 min in a small (7.5 × 7.5 × 15 cm) plastic cage. The tester used an Austerlitz pin (000; Fine Scientific Tools) to touch the plantar surface of the hindpaw without skin penetration. Numbers of withdrawal responses per 10 trials (within 1 min) were counted.

Tape removal. Mice were habituated in a transparent cylinder. A 9-mm diameter circular adhesive Microtube Tough-Spots label (Diversified Biotech) was attached to the plantar surface of the hindpaw. The latency to biting or licking to remove the tape (sense time) was measured. Three measurements were performed per mouse.

Brush (dynamic mechanical stimuli). Each mouse was habituated in a small (7.5 × 7.5 × 15 cm) plastic cage for at least 20 min before testing. The plantar hindpaw was stimulated by light stroking from heel to toe with a paintbrush. For baseline dynamic mechanical sensitivity, the test was repeated ten times. The number of occurrences of walking away or occasionally brief paw lifting was measured. For dynamic mechanical allodynia, the paw flinching/licking time was recorded as an indicator and averaged over three trials for individual animals.

von Frey filaments (punctate mechanical stimuli). Each mouse was habituated in a small (7.5 × 7.5 × 15 cm) plastic cage for at least 20 min before testing. Mechanical sensitivity was determined with a series of von Frey filaments (bending forces: 0.02, 0.04, 0.07, 0.16, 0.4, 0.6, 1 and 2 g) applied within the sciatic nerve territory (lateral part of the hindpaw). Each filament was tested ten times in increasing order from the lowest force. Between individual measurements, von Frey filaments were applied at least 3 s after the mice had returned to their initial resting state. For baseline static mechanical sensitivity, all filaments were applied and the number of withdrawals was recorded. For tactile allodynia, the minimal force filament for which animals presented either a brisk paw withdrawal and/or an escape attempt in response to at least five of the ten stimulations was determined as the mechanical response threshold.

Open field. To test overall locomotor activity, mice were individually placed on the side of a circular open arena that consisted of an unmarked floor enclosed by a 45 cm diameter wall and allowed to roam freely for 20 min. The arena was divided into the outermost periphery (wall), the centre of the arena (15 cm diameter), and the neutral zone between the centre and the wall (25 cm diameter). Video tracking

© 2018 Springer Nature Limited. All rights reserved.

software (Ethovision XT v11.5, Noldus BV) was used to quantify the distance travelled in each zone by the animal. The arena was housed in a dimly lit room (30 lx) with no noise, and examiners were not present in the room during testing, to ensure limited distractions to the mice.

Ground walking. For ground walking, mice were placed in the MotoRater (TSE Systems)[12] and kinematic analysis was performed based on data collected by the MotoRater.

Rotarod. Mice were trained five times on a single speed (15 r.p.m.) rotarod (IITC Life Science). After that, the latency to fall was recorded on the next day. The experiment was repeated twice at 20-min intervals and the average was calculated as rotarod latency.

Calcium imaging in freely behaving mice. Mice used for calcium imaging experiments were injected with HiRet-Cre (5×10^{12} copies/ml) into the spinal cord on P14. In adult mice, a cranial window was installed over the injected area as described[12,31–33]. During the procedures, mice were anaesthetized with Avertin (1.5% solution given at 0.01 ml/g, i.p.) and treated with dexamethasone (0.2 mg/kg, s.c.) and carprofen (5 mg/kg, s.c.) to prevent brain swelling and inflammation. A piece of skull (3.5 mm in diameter) above the somatosensory cortex was removed after high-speed dental drilling. AAV2/9-FLEX-syn-GCaMP6 s (5×10^{13} copies/ml, 0.45 μl per animal) was injected into primary sensory (hindlimb S1) cortex using a micro-syringe pump. A 3-mm coverslip was used to seal the cranial window and the exposed scalp was sutured. After 7 days, the baseplate of a miniaturized integrated fluorescent microscope (Inscopix)[34] was fixed on top of the coverslip[17]. Animals were then habituated to the attachment of the microscope for 2 days. Calcium imaging was then performed in freely moving mice without sensory stimulation, or in intact mice with light brush, von Frey fibre and laser heat applied to the hindpaw. SNI was then performed and calcium imaging was conducted on the same animals at 3 days post-SNI. Sham controls were conducted in separate animals by cutting the skin on the hindlimb without damaging the nerves involved in SNI, and imaging sessions were performed before and after sham procedures. Each imaging session include 5 blocks: 20 min habituation, 4 min free moving followed by 13–17 trials (trial interval: 10 s) of brush, von Frey and laser heat stimuli.

Calcium imaging was performed in freely moving mice using the head-attached microscope (Inscopix; LED power: 0.6–1.0 mW; camera resolution: 1,440 × 1,080 pixels). Images were acquired at 30 Hz using nVista HD software (Inscopix). At the beginning of each imaging session, the protective cap of the previously implanted baseplate was removed and the microscope was attached. The imaging field of view was approximately $900 \times 600\,\mu m^2$ at 0.65 μm/pixel resolution and the imaging depth was selected by adjusting the focus of the microscope until clear dendritic trunk signals were observed, which appeared as bright spots in the images. The focal plane was 150–250 μm below the lens. Mouse behaviour was recorded with a video camera (Canon), which was synchronized with calcium imaging using the trigger-out signal from nVista HD[12].

Calcium imaging videos were analysed using Mosaic software (Inscopix) and custom-written scripts in MATLAB following published algorithms[35,36]. Raw videos were first down-sampled by fourfold along spatial dimensions to reduce file size and noise. The mean fluorescence intensity of each pixel during a recording session (4–8 min) was calculated as F_0 and changes in pixel intensity at time t were expressed as $(F_t - F_0)/F_0$ or $\Delta F/F_0$. To extract active dendritic signals, principal component and independent component analysis (PCA-ICA) was applied to the spatial-temporal data matrices of $\Delta F/F_0$ using CellSort and fastICA toolboxes (these toolboxes are freely downloadable from MATLAB Central[35,36]). This analysis decomposes a spatiotemporal data matrix into independent components based on the skewness of data distribution. Each component has a characteristic spatial filter over the imaged area and a corresponding temporal signal during the imaging period. The spatial filter and the temporal signal of each component were graphically displayed and inspected by human observers who were blind to the experimental conditions of each video. If the spatial filter for a component overlapped with the dark shadows casted by blood vessels in the F_0 image, this component was likely to be contributed by blood flow and was therefore rejected. In addition, because calcium signals show characteristic fast-rising and slow-decaying time course[37], the temporal skewness of calcium signals is expected to be positive and those components with skewness less than 1 were rejected[35]. For each selected component, the location of the dendritic trunk was identified as the brightest spot (3 × 3 pixel) of the spatial filter. The corresponding temporal signal of the dendritic trunk was calculated from the $\Delta F/F_0$ video by subtracting the median value of the background area (outside the dendritic tree) from the average value of the dendritic trunk area. Calcium activity in dendritic trunks tightly correlated with that in the soma in freely behaving mice[12], similar to what has been reported before by two-photon microscopy[38,39].

To identify periods of increased neuronal activity, we searched for the rising phase of each calcium event (peak $\Delta F/F_0 > 3$ s.d. of baseline fluctuation), which has been shown to be closely associated with neuronal spiking activity[37,40,41]. The start of this rising phase is detected when the first derivative of $\Delta F/F_0$ (calculated in a 200-ms moving window) rises above 0 and continues to increase above 5 s.d. of baseline fluctuation[41], and the end of this rising phase is detected when the first derivative of $\Delta F/F_0$ falls below 0. For each dendrite, the amplitude of each detected calcium event was normalized by the s.d. of the whole calcium trace. To visualize the activity patterns of detected dendrites during stimulation, the active event traces of each CSN were aligned to the time when the brush, von Frey filament or laser heat was applied to the left plantar surface of the hindlimb. The resulting traces from all the CSNs were sorted by their peak activation time during the window and displayed in temporal raster plots.

To quantify the difference in activity between pre- and post-SNI conditions, we first calculated the trial-averaged activities of CSN dendrites in response to different sensory stimuli, then compared the responses between pre- and post-SNI conditions within 1 s of stimulation (rank sum tests). A significant increase in CSN activity in response to von Frey filaments was observed after SNI (Fig. 3c), but not in sham controls (data not shown).

To quantify the proportion of activated neurons responding to brush only, von Frey only, or both stimuli in pre- and post-SNI conditions, we concatenated the calcium videos of the brush and von Frey sessions. Responsive neurons were classified based on whether they showed maximal trial average activity within 1 s after either or both stimuli.

Spinal cord slice preparation and electrophysiology. To label CST axons, AAV-2/1-ChR2-YFP was injected unilaterally into the somatosensory cortex in neonatal CCK–tdTomato mice (P1). At P28–P31, mice were deeply anaesthetized with isoflurane, decapitated and the lumbar spinal cord was quickly removed to ice-cold modified artificial cerebrospinal fluid (ACSF): NaCl, 80; KCl, 2.5; NaH_2PO_4, 1.25; $CaCl_2$, 0.5; $MgCl_2$, 3.5; $NaHCO_3$, 25; sucrose, 75; sodium ascorbate, 1.3; sodium pyruvate, 3.0 (all in mM), with pH at 7.4 and osmolarity at 310–320 mOsm, and oxygenated with 95% O_2 and 5% CO_2. Sagittal spinal cord slices (350–500 μm) with dorsal roots (3–5 mm) attached were cut using a vibratome (VT1000S, Leica). Slices were then incubated for about 1 h at 35 °C in a solution that contains (in mM): NaCl, 125; KCl, 2.5; $CaCl_2$, 2; $MgCl_2$, 1, NaH_2PO_4, 1.25; $NaHCO_3$, 26; D-glucose, 25; sodium ascorbate, 1.3; sodium pyruvate, 3.0 with pH at 7.2 and measured osmolality at 310–320 mOsm, and oxygenated with 95% O_2 and 5% CO_2. The slice was then transferred into a recording chamber and perfused with oxygenated recording solution at a rate of 5 ml/min before electrophysiological recordings at room temperature (23 °C).

For whole cell patch-clamp recording, CCK–tdTomato neurons were viewed using an upright microscope (Olympus, BX51WI) with a 63× water-immersion objective, infrared differential interference contrast (IR-DIC) and fluorescence. Patch pipettes (5–10 MΩ) were made from borosilicate glass (Sutter) using a P-1000 horizontal micropipette puller (Sutter). Internal solution contained (in mM): potassium gluconate 130, KCl 5, MgATP 4, NaGTP 0.5, HEPES 10, EGTA 0.5, pH 7.28 with KOH, and measured osmolality at 280–300 mOsm. This internal solution allows separation of AMPAR and $GABA_A$ currents at the reversal potential of the other receptor. Data were acquired with pClamp 10.0 software (Molecular Devices) using a MultiClamp 700B patch-clamp amplifier and Digidata 1440A (Molecular Devices). Responses were low-pass filtered online at 4 kHz and digitized at 10 kHz. Junction potential was around 6–8 mV. No correction for junction potential was applied. Membrane capacitance (24.3 ± 1.1 pF), membrane resistance (493.4 ± 29.4 mΩ) were monitored continuously and cells were excluded from analysis if these values changed by more than 10% during an experiment. Following established procedure[21], recordings with series resistance greater than 30 mΩ or resting membrane potential more depolarized than −50 mV were excluded from analysis. All recordings were performed at room temperature (23 °C).

Voltage clamp recordings were performed by holding membrane potential at −70 mV. Under this condition, evoked inhibitory postsynaptic currents (eIPSCs) were minimized, and evoked excitatory postsynaptic currents (eEPSCs) were detected. This was used to study whether a neuron received inputs directly (monosynaptic) or indirectly (polysynaptic) from dorsal root or CST axons. Current clamp was also carried out at resting membrane potential in order to record evoked action potential output. Dorsal root stimulation was generated with ISO-Flex stimulus isolator (A.M.P.I.) and delivered with suction electrode (A-M Systems). The stimulation pulse width was 0.1 ms and intensity was 25 μA for Aβ fibres. High-frequency stimulation at 20 Hz was used to determine monosynaptic input for Aβ and CST fibres. Laser stimulation (473 nm, 5 ms, 5 mW) for activating CST axon terminals was delivered using an optical fibre (200 μm core, 0.22NA; Thorlabs) directly illuminating the recording area. Monosynaptic transmission for opto-stimulation was verified by applying tetrodotoxin (TTX; 0.5 μM, Abcam) in the recording bath to block sodium channels, followed by 4-aminopyridine (4-AP; 2 mM, Sigma) to augment light-induced, direct depolarization of ChR2 axon terminals[42]. NBQX (5 μM) and CPP (20 μM) were used to block AMPAR and NMDAR currents, respectively. To mimic the spinal disinhibition condition, both

© 2018 Springer Nature Limited. All rights reserved.

voltage and current clamps were performed in the presence of both bicuculline (10 μM, Sigma) and strychnine (2 μM, Sigma).

Innocuous mechanical stimulation for c-Fos and/or pERK induction. To induce c-Fos expression by innocuous mechanical (dynamic brush) stimulation, animals were lightly restrained using a rodent restraint bag (traditional triangle shape, AIMS). We used a paintbrush to stroke the hindpaw. Each touch, lasting 2 s and moving from the middle of the foot to the distal foot pad along the peripheral side that belongs to the sural nerve territory, was applied once every 4 s for 10 min. This touch stimulus does not elicit a flexion reflex in normal mice. Animals were perfused two hours later and spinal cord transverse sections were prepared for c-Fos immunostaining. For pERK induction, the same procedures were applied, except that mice were killed immediately after repetitive mechanical brushing or 5 min after thermal stimulation.

Light-based CSN stimulation. The HiRet-Cre virus (2×10^{12} copies/ml) was injected intraspinally into the lumbar spinal cord of LSL-ChR2-YFP mice at P14 to retrogradely label hindlimb corticospinal neurons. Four weeks later, the animals were subjected to optical stimulation experiments. For optic fibre implantation, animals were anaesthetized with ketamine/xylazine (100 mg/kg ketamine, 10 mg/kg xylazine) and mounted on a stereotaxic frame. A 1-mm hole was drilled on the right side of the skull, centred at −1.2 mm posterior to Bregma and 1.65 mm lateral to midline, with the dura left intact. All SNI surgeries were performed on the left hindpaw. A customized fibre optic cannula (Ø1.25 mm Ceramic Ferrule, Ø200 μm Core, 0.50 NA, 2 mm projection; Thorlabs) was positioned on the cortical surface through the hole. Silicon elastomer and dental cement were applied to protect the skull window and secure the cannula. The incision was sutured and animals were allowed to recover for 7 days before behavioural testing. Before optogenetic stimulation, an optogenetic patch cable was connected to the cannula via a mating sleeve. The patch cable was connected to the laser source with a rotary joint (FRJ 1×1, Doric) to allow free movement of the animal. Laser stimulation (473 nm, 5mW; 100 pulses of 5 ms, 20 Hz) was delivered during sensory tests. We verified the expression of ChR2-YFP in the hindlimb S1 by post hoc histological analysis.

For pERK induction, animals were lightly restrained using a rodent restraint bag (traditional triangle shape, AIMS). Laser stimulation (473 nm, 5 mW; 100 pulses of 5 ms, 20 Hz) was delivered at 2-s intervals for 10 min, and repeated three times with a 1-min rest between each session in mice with SNI (contralateral cortex). Repetitive light brush (see 'Innocuous mechanical stimulation for c-Fos and/or pERK induction') was applied during optogenetic stimulation. Animals were perfused immediately after stimulation.

Quantification and statistical analysis. The normality and variance similarity were measured by STATA (version 12) before we applied any parametric tests. Two-sided or one-sided *t*-tests were used for single comparisons between two groups. The rest of the data were analysed using one-way or two-way ANOVAs, depending on the appropriate design. Post hoc comparisons were carried out only when a main effect showed statistical significance. *P* values of multiple comparisons were adjusted using Bonferroni's correction. Error bars in all figures represent mean ± s.e.m. $***P < 0.001$, $**P < 0.01$, $*P < 0.05$, NS, no statistical significance ($P \geq 0.05$).

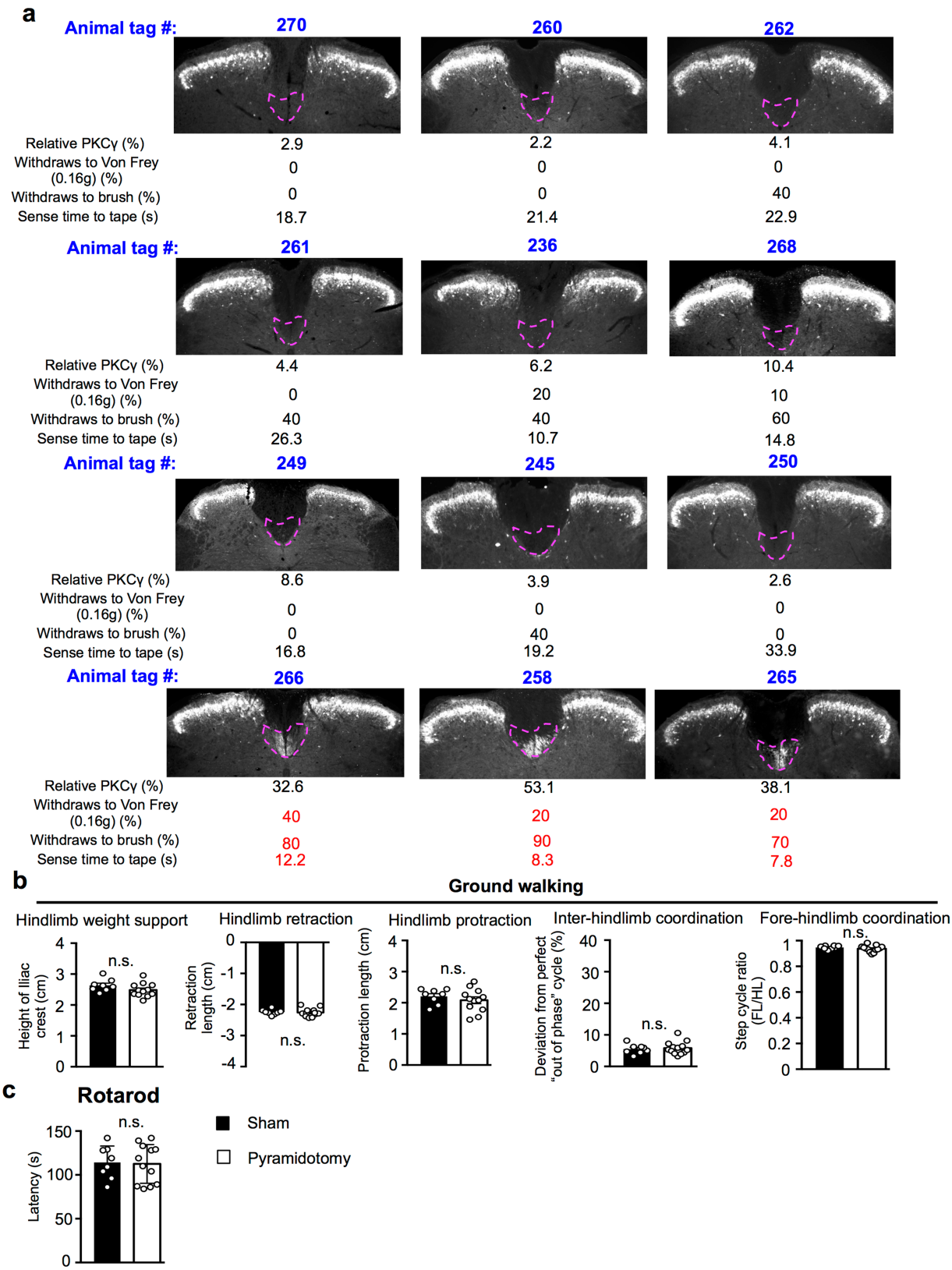
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. All custom codes used in this study are available from the corresponding author on reasonable request.

Data availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

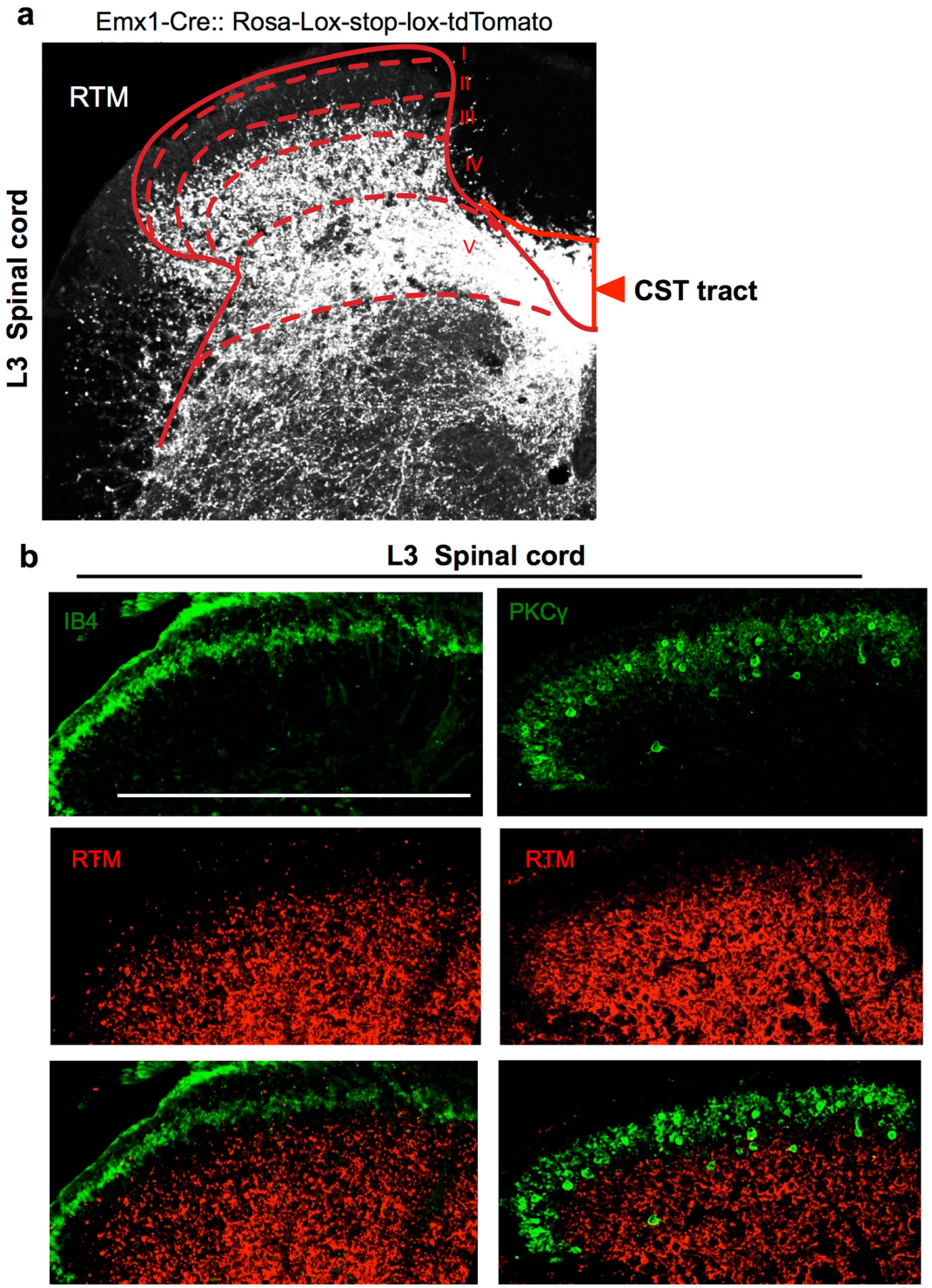
27. Sun, F. et al. Sustained axon regeneration induced by co-deletion of PTEN and SOCS3. *Nature* **480**, 372–375 (2011).
28. Arlotta, P. et al. Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. *Neuron* **45**, 207–221 (2005).
29. Metz, G. A., Dietz, V., Schwab, M. E. & van de Meent, H. The effects of unilateral pyramidal tract section on hindlimb motor performance in the rat. *Behav. Brain Res.* **96**, 37–46 (1998).
30. Muir, G. D. & Whishaw, I. Q. Complete locomotor recovery following corticospinal tract lesions: measurement of ground reaction forces during overground locomotion in rats. *Behav. Brain Res.* **103**, 45–53 (1999).
31. Mastwal, S. et al. Phasic dopamine neuron activity elicits unique mesofrontal plasticity in adolescence. *J. Neurosci.* **34**, 9484–9496 (2014).
32. Cao, V. Y. et al. In vivo two-photon imaging of experience-dependent molecular changes in cortical neurons. *J. Vis. Exp.* **71**, 50148 (2013).
33. Cao, V. Y. et al. Motor learning consolidates Arc-expressing neuronal ensembles in secondary motor cortex. *Neuron* **86**, 1385–1392 (2015).
34. Ziv, Y. et al. Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264–266 (2013).
35. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63**, 747–760 (2009).
36. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430 (2000).
37. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
38. Hill, D. N., Varga, Z., Jia, H., Sakmann, B. & Konnerth, A. Multibranch activity in basal and tuft dendrites during firing of layer 5 cortical neurons in vivo. *Proc. Natl Acad. Sci. USA* **110**, 13618–13623 (2013).
39. Peters, A. J., Lee, J., Hedrick, N. G., O'Neil, K. & Komiyama, T. Reorganization of corticospinal output during motor learning. *Nat. Neurosci.* **20**, 1133–1141 (2017).
40. Cichon, J., Blanck, T. J. J., Gan, W. B. & Yang, G. Activation of cortical somatostatin interneurons prevents the development of neuropathic pain. *Nat. Neurosci.* **20**, 1122–1132 (2017).
41. Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).
42. Holloway, B. B. et al. Monosynaptic glutamatergic activation of locus coeruleus and other lower brainstem noradrenergic neurons by the C1 cells in mice. *J. Neurosci.* **33**, 18792–18805 (2013).

© 2018 Springer Nature Limited. All rights reserved.



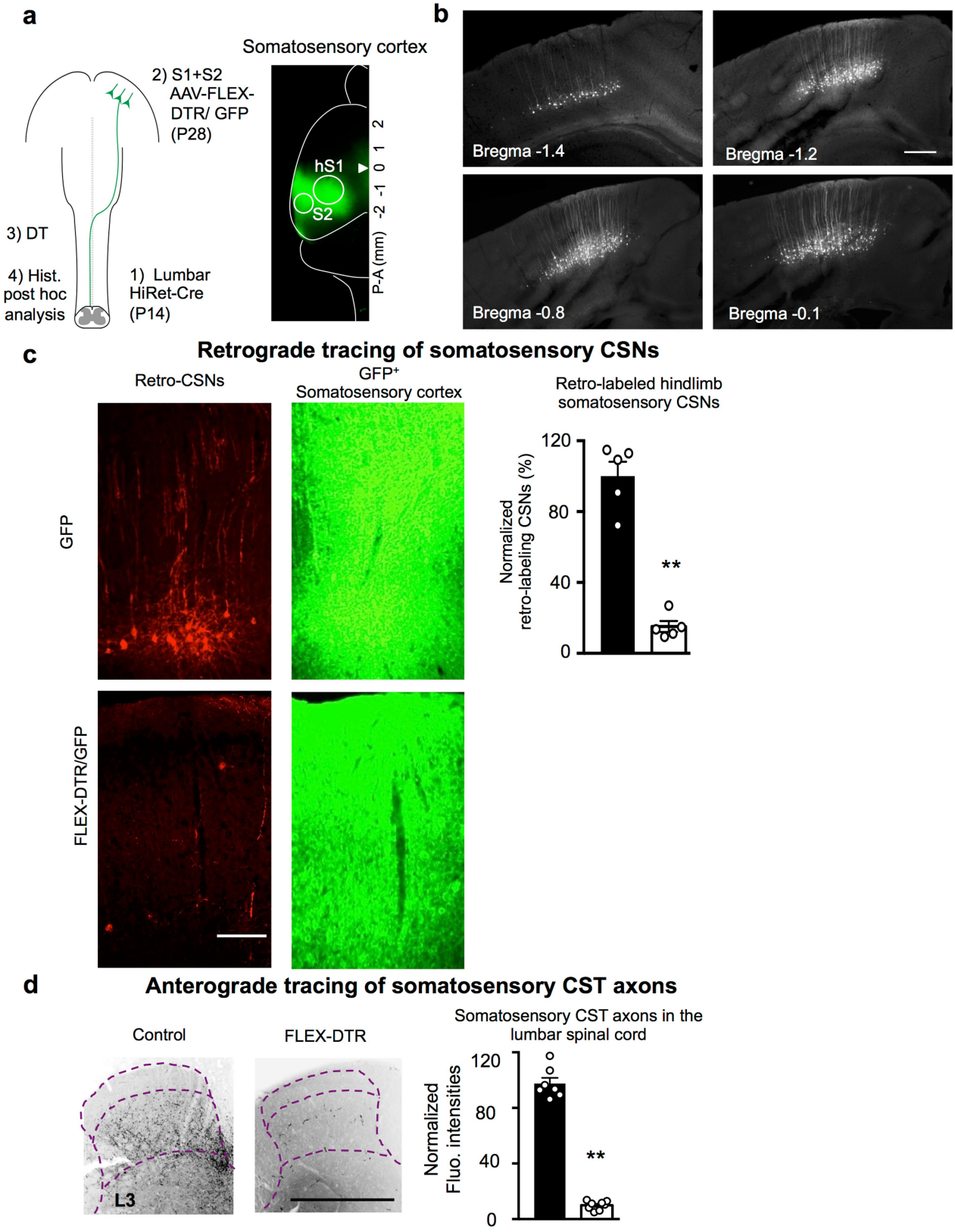
Extended Data Fig. 1 | Effects of pyramidotomy on tactile behaviour and gross locomotion in mice. a, Correlation between CST ablation and tactile behaviours in mice with pyramidotomy. For individual animals that received pyramidotomy, tag number, image and quantification of L3 spinal cord sections stained with anti-PKCγ antibodies showing remaining CST axons, percentage of withdrawal response to low-threshold von Frey filament (0.16g) and light brush, and sense time to the tape are shown. **b**, **c**, Performance on ground walking (**b**, $P = 0.14$, hindlimb weight support; $P = 0.81$, hindlimb retraction; $P = 0.41$, hindlimb protraction; $P = 0.81$, inter-limb coordination; $P = 0.20$, fore–hindlimb coordination), and rotarod test (**c**, $P = 0.87$) in mice with sham surgery ($n = 8$) or pyramidotomy ($n = 12$). n.s., no statistical significance. Two-sided student's t-test. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | CST axon termination in the lumbar spinal cord. **a**, Representative transverse spinal section (L3) from an Emx1-tdTomato (red) reporter line. **b**, Sections were co-stained with IB4 (green), a lamina IIi marker, and anti-PKCγ, a laminae IIi/III marker, in the spinal dorsal horn. Scale bar, 500 μm. For **a** and **b**, three and four mice, respectively, showed similar results.

© 2018 Springer Nature Limited. All rights reserved.

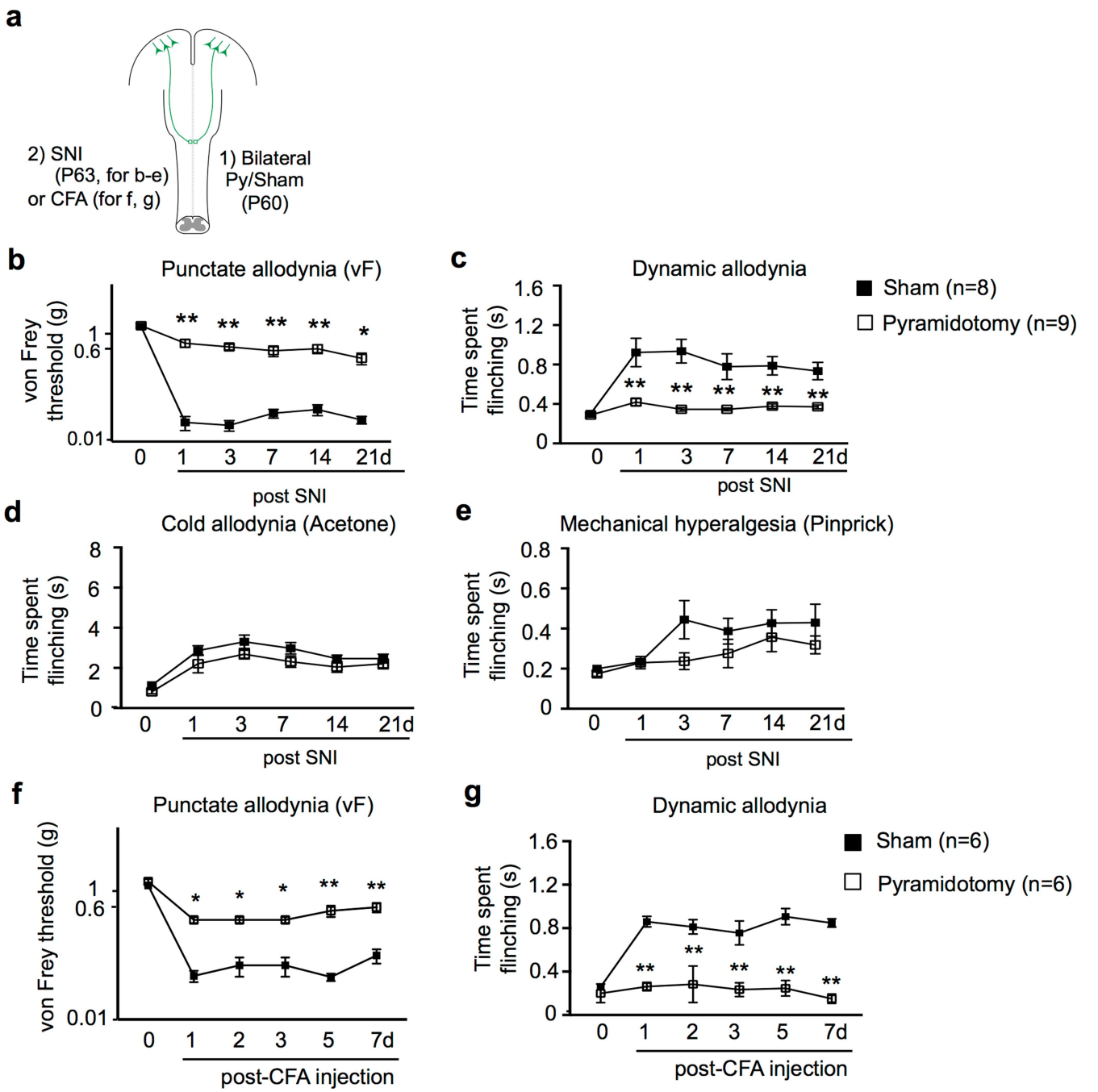


Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

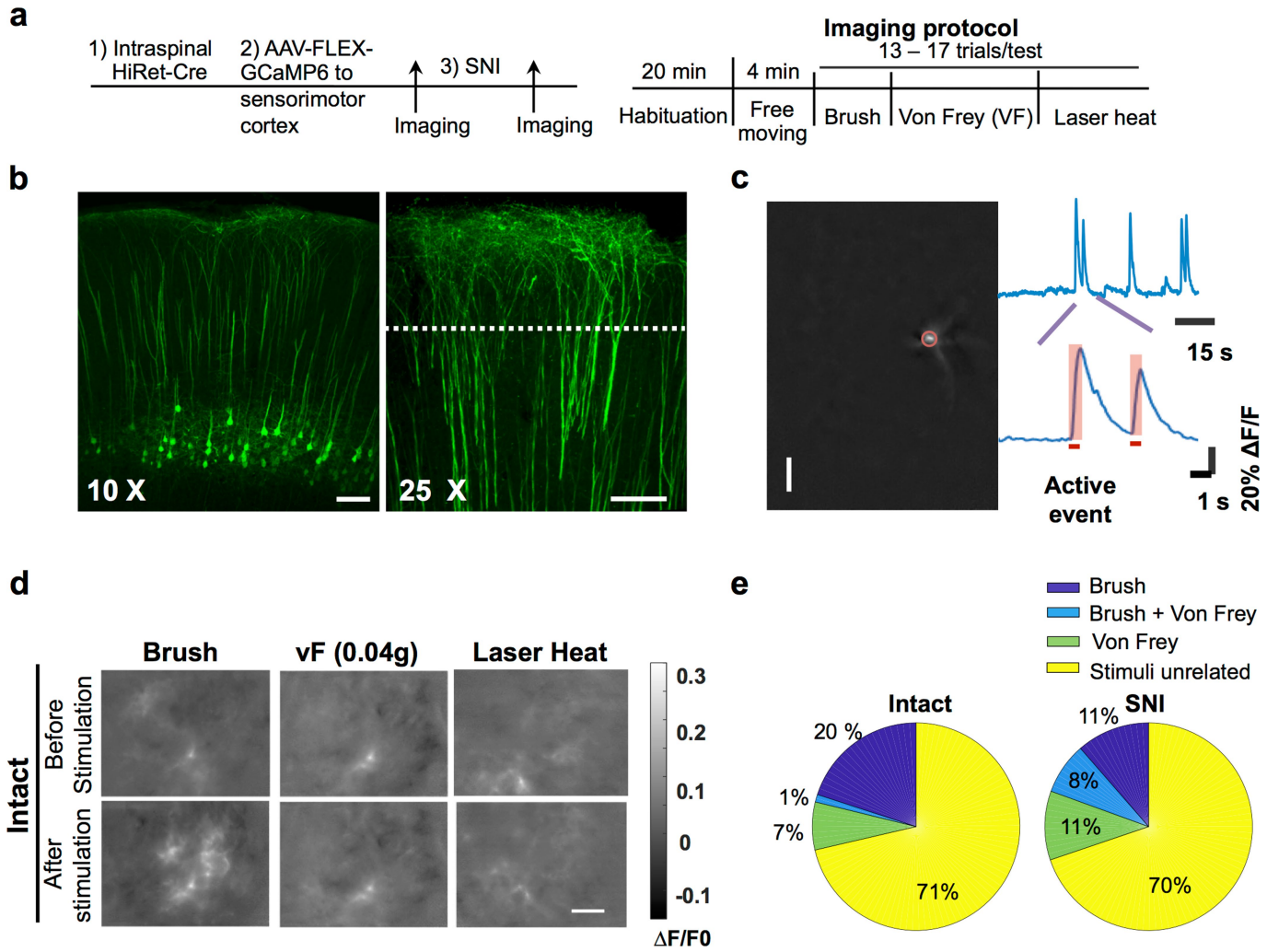
Extended Data Fig. 3 | Efficiency of somatosensory CSN ablation by P14 intraspinal injection. a, Left, schematic of regional CSN ablation by P14 lumbar (T13–L6) intraspinal injection. Right, representative image ($n = 8$ animals with similar results) of the cortex with GFP^+ areas covering hindlimb S1/S2. **b**, Representative images ($n = 6$ animals with similar results) of cortical sections showing hindlimb CSNs retrogradely labelled by lumbar (T13–L6) intraspinal injection of HiRet-GFP at P14. Scale bar, 100 μm. **c**, To assess ablation efficiency, at the end point, retrograde-targeting rAAV-mCherry was injected into the lumbar spinal cord in some animals. Representative images of cortical sections showing retrogradely labelled $mCherry^+$ CSNs (left) within the GFP^+ cortical areas (right) (S1/S2) in control or AAV-FLEX-DTR injected animals with quantification (normalized to those in controls as 100). $^{**}P < 0.01$ ($P < 0.0001$), two-sided Student's *t*-test. $n = 5$ and 5 for control and AAV-FLEX-DTR injected mice, respectively. Scale bar, 100 μm. **d**, Representative images of transverse lumbar spinal cord sections showing residual CST axons labelled by GFP (from mice co-injected with AAV-GFP to S1/S2 and AAV-FLEX-DTR) in control or S1/S2 CSN ablated animals with quantification. $^{**}P < 0.01$ ($P < 0.0001$), two-sided Student's *t*-test. $n = 7$ or 8 for control or AAV-FLEX-DTR-injected mice, respectively. Scale bar, 500 μm. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



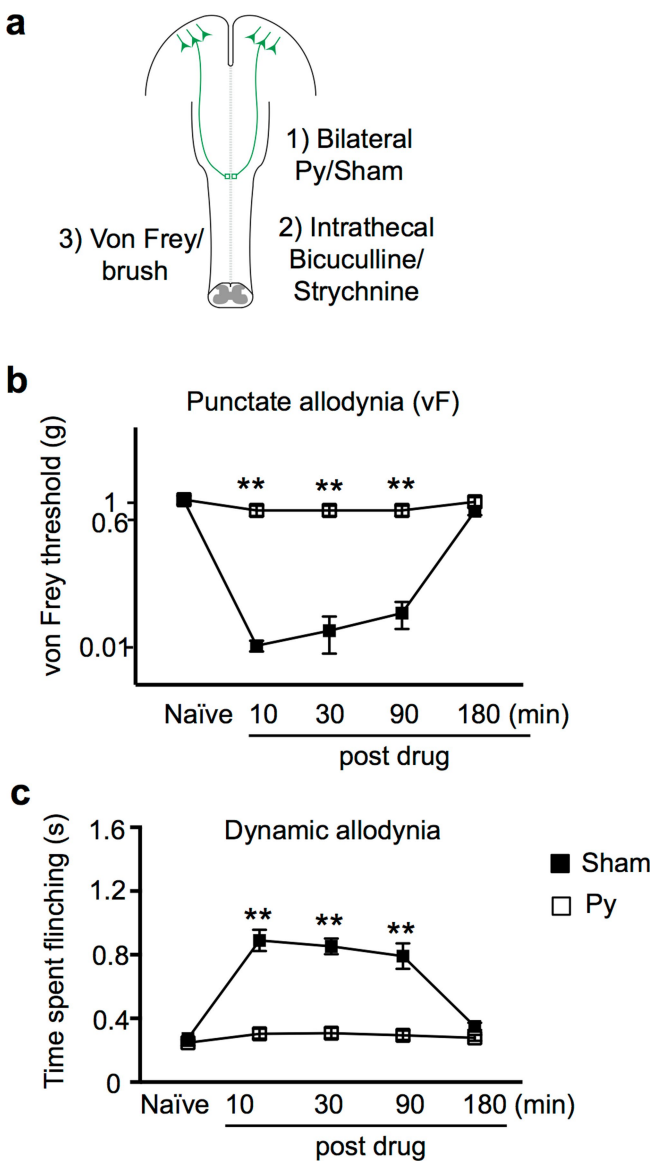
Extended Data Fig. 4 | Mechanical allodynia induced by SNI or CFA injection is compromised in mice with pyramidotomy, but cold allodynia and mechanical hyperalgesia induced by SNI are not. **a**, Schematic drawing of experimental paradigm. **b**–**e**, Measurement of punctate (**b**) and dynamic (**c**) mechanical allodynia, cold allodynia (**d**) and mechanical hyperalgesia (**e**) after SNI in mice that underwent sham surgery ($n = 8$) or pyramidotomy ($n = 9$) 1–21 days after SNI. **b**, $P < 0.0001$, $P < 0.0001$, $P = 0.0012$, $P = 0.0004$ and $P = 0.041$; **c**, $P < 0.0001$, $P < 0.0001$, $P = 0.0004$, $P = 0.001$ and $P = 0.0045$; **d**, $P = 0.26$, $P = 0.33$, $P = 0.29$, $P > 0.99$ and $P > 0.99$; **e**, $P > 0.99$, $P = 0.15$, $P = 0.56$, $P > 0.99$ and $P = 0.74$ for 1, 3, 7, 14 and 21 d post SNI, respectively. **f**, **g**, Measurement of punctate (**f**) and dynamic (**g**) mechanical allodynia in mice that underwent sham surgery ($n = 6$) or pyramidotomy ($n = 6$) 1–7 days after hindpaw CFA injection. **f**, $P = 0.01$, $P = 0.01$, $P = 0.01$, $P < 0.0001$ and $P < 0.0001$; **g**, $P < 0.0001$ for 1, 2, 3, 5 and 7d post CFA injection, respectively. Two-way repeated measures ANOVA followed by Bonferroni correction. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



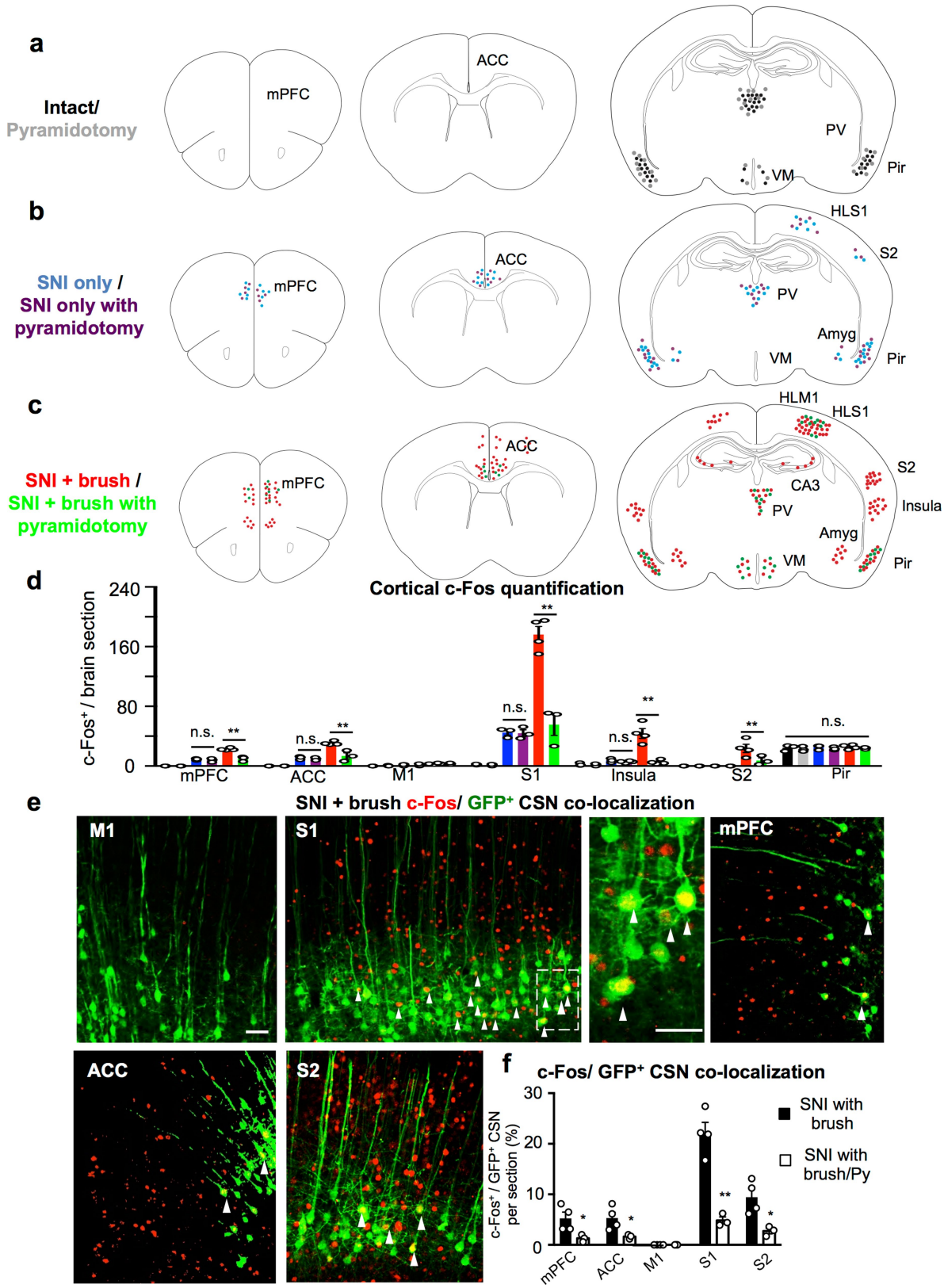
Extended Data Fig. 5 | Calcium imaging of CSN activity in intact and SNI mice. a, Schematic of experimental procedures. **b**, Confocal fluorescence images of coronal brain sections showing specific expression of GCaMP6s in CSNs. Left, a 10× image showing labelled CSN soma and dendrites; right, a 25× image showing the magnified view of apical dendritic trunks. Dotted line, expected focal plane of head-mounted microscope. Scale bars, 100 μm. **c**, Procedures for identifying the active events of CSN dendrites. Left, example of dendrites identified from a calcium movie by ICA analysis. The brightest spot in a dendritic tree, corresponding to the trunk (red circle), is used as a region of interest for temporal signal analysis. Top trace, temporal signal of the dendrite. Bottom trace, magnified calcium events. Horizontal bars indicate rising phases of calcium events, which are associated with neuronal activation and used in subsequent analysis. Scale bar, 100 μm. **d**, Example calcium movie frames showing dendritic activities of hindlimb S1 CSNs upon different sensory stimuli in intact mice. In these examples, brush stimuli activated CSNs, whereas von Frey (0.04 g) and laser heat stimuli did not. Calcium signals are expressed as $\Delta F/F_0$ (F_0 is the time-average fluorescence of the whole movie). Scale bar, 200 μm. For **b**, **d**, the experiments were repeated independently four times with similar results. **e**, Pie charts showing the proportions of neurons that responded to brushes, von Frey filaments or both brush and von Frey stimulation before and after SNI. Few neurons responded to both stimuli before SNI, but this overlapping proportion increased after SNI.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Mechanical allodynia induced by spinal disinhibition (treatment with bicuculline and strychnine) is compromised in mice with pyramidotomy. a, Drawing of experimental paradigm. **b**, **c**, Measurement of punctate (**b**) and dynamic (**c**) mechanical allodynia after intrathecal injection of bicuculline and strychnine in mice that had undergone sham surgery ($n = 6$) or pyramidotomy ($n = 7$). $^{**}P < 0.01$ ($P < 0.0001$ for **b** and **c** at 10, 30, and 90 min post drug), two-way repeated measures ANOVA followed by Bonferroni correction. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

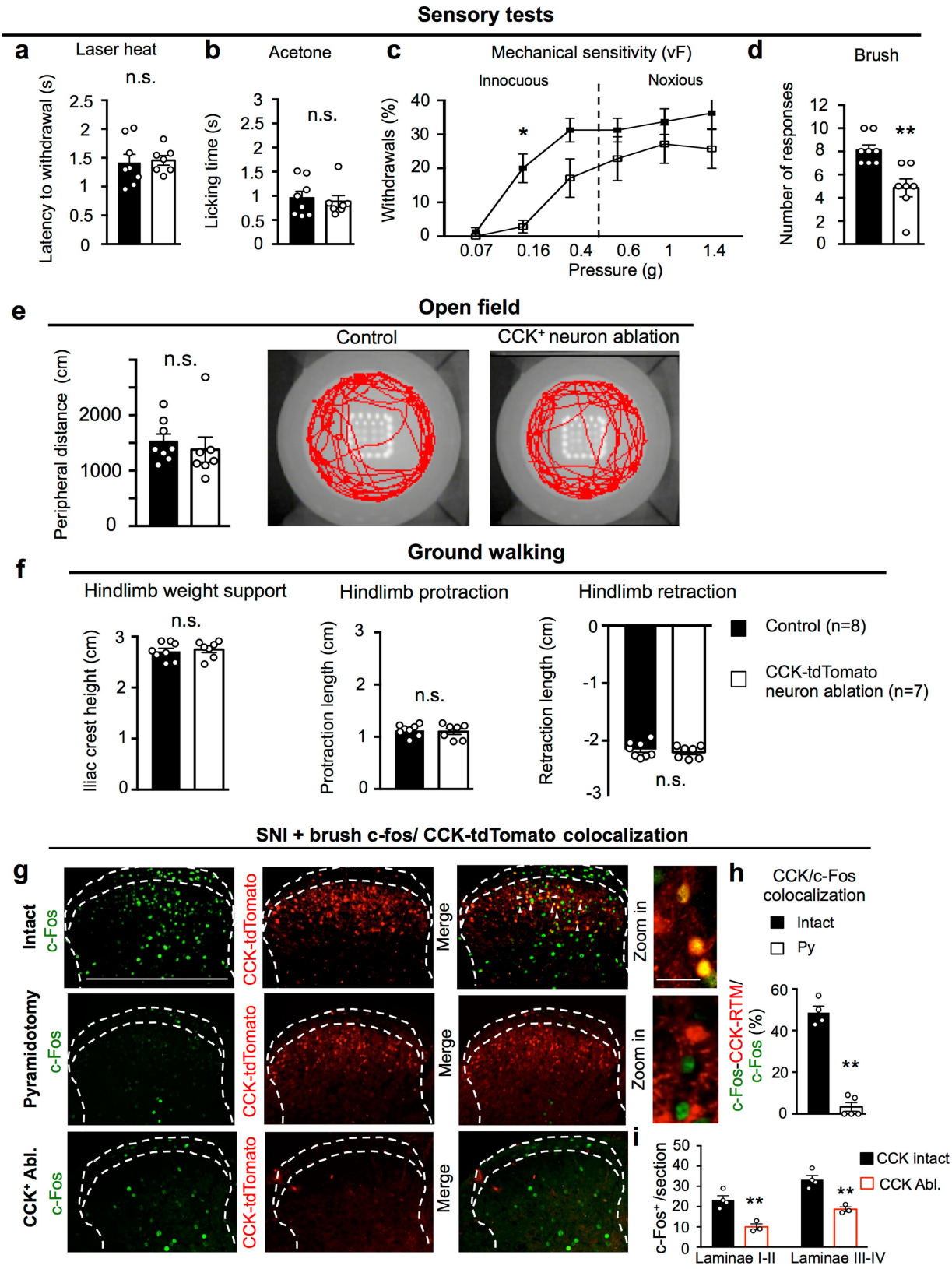


Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 7 | Neuronal activity in cortical and subcortical areas upon light touch after SNI. a–c, Drawings of c-Fos immunostaining in intact mice (**a**), mice with SNI only (**b**) and mice with SNI after light brush stimulation (**c**) from control mice and mice with pyramidotomy. mPFC, medial prefrontal cortex; ACC, anterior cingulate cortex; S1HL, hindlimb primary somatosensory cortex; S2, secondary somatosensory cortex; Pir, piriform cortex; PV, periventricular nucleus of the thalamus; VM, ventromedial nucleus of the hypothalamus; Amyg, amygdala. **d**, Quantification of c-Fos$^+$ cells in multiple cortical areas in intact mice (with CSN, $n = 3$; with pyramidotomy, $n = 3$), mice with SNI only (with CSN, $n = 3$; with pyramidotomy, $n = 3$), and mice with SNI after light brush stimulation (with CSN, $n = 4$; with pyramidotomy, $n = 3$). $^{**}P < 0.01$; n.s., no statistical significance. PFC SNI only with or without Py: $P > 0.99$; PFC SNI + brush with or without Py: $P < 0.0001$; ACC SNI only with or without Py: $P > 0.99$; ACC SNI + brush with or without Py: $P = 0.0002$; M1 SNI only with or without Py: $P > 0.99$; M1 SNI + brush with or without Py: $P > 0.99$; S1 SNI only with or without Py: $P > 0.99$; S1 SNI + brush with or without Py: $P < 0.0001$; insula SNI only with or without Py: $P > 0.99$; insula SNI + brush with or without Py: $P < 0.0001$; S2 SNI + brush with or without Py: $P = 0.0075$; Pir all conditions (ANOVA): $P = 0.82$. One-way ANOVA followed by Bonferroni correction. **e**, Representative images of multiple cortical areas stained with c-Fos (red) and GFP (green, HiRet-GFP injection) in mice with SNI after light brush stimulation. Arrowheads mark the co-localization of c-Fos$^+$ and GFP$^+$ CSNs. Scale bars, 20 μm. **f**, Quantification of c-Fos$^+$ and GFP$^+$ CSN co-localization in multiple cortical areas in animals with SNI after light brush stimulation without ($n = 4$) or with pyramidotomy (Py, $n = 3$). $^{*}P < 0.05$; $^{**}P < 0.01$, $P = 0.03$, 0.04, 0.0009, and 0.02 for mPFC, ACC, S1, and S2, respectively. Two-sided Student's *t*-test. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

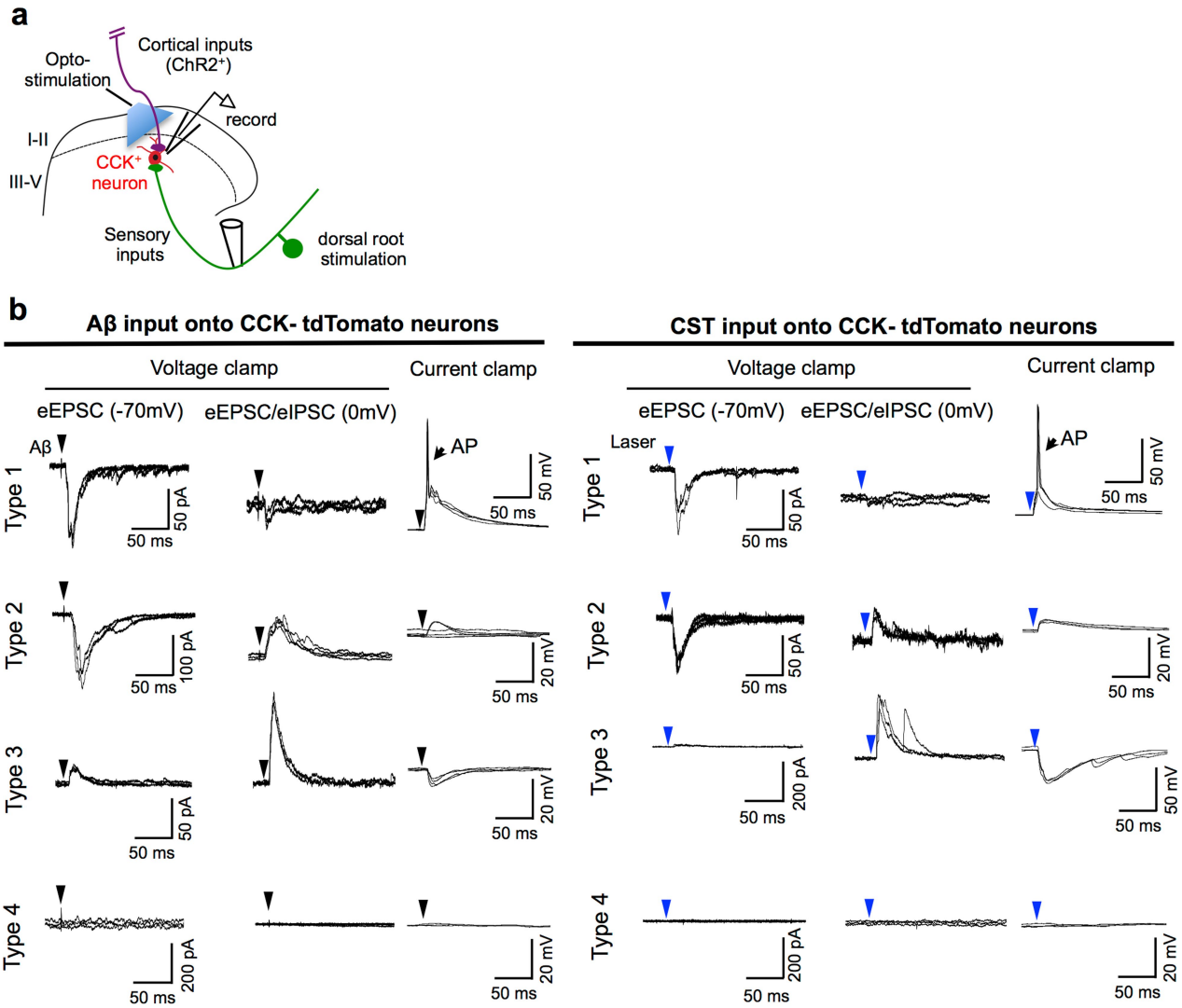


Extended Data Fig. 8 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 8 | Ablation of lumbar CCK+ interneurons reduces tactile sensitivity and dorsal horn neuronal activation after SNI, but not nociceptive response or gross locomotion. a–**d**, Measurements of sensitivity to laser heat (**a**, $P = 0.76$), acetone (**b**, $P = 0.86$), von Frey filaments (**c**, $P = 0.03$ for 0.016 g) and brush (**d**, $P = 0.002$) stimuli in control mice ($n = 8$) or mice in which CCK–tdTomato interneurons had been ablated ($n = 7$). **a**, **b**, **d**, Two-sided Student's *t*-test; **c**, two-way repeated measures ANOVA followed by Bonferroni correction. **e**, **f**, Performance on open field (**e**, $P = 0.54$) and ground walking (**f**, $P = 0.68$, 0.72, and 0.50 for hindlimb weight support, protraction, and retraction, respectively) in control mice ($n = 8$) or mice in which CCK–tdTomato interneurons had been ablated ($n = 7$). n.s., no statistical significance, two-sided Student's *t*-test. **g**–**i**, Representative images of c-Fos (green) activity (**g**) and quantification of CCK^+/c-Fos^+ cells (**h**) and c-Fos^+ neurons in different laminae of the dorsal horn of the spinal cord (L3–4) of CCK–tdTomato (red) mice after SNI and brush stimulation in mice that had undergone sham surgery ($n = 4$), pyramidotomy ($n = 5$) or lumbar CCK–tdTomato ablation ($n = 3$) (**i**). Scale bars, 500 μm (**g**, right, middle, and left columns) and 50 μm (**g**, zoom in panels). $^{**}P < 0.01$, two-sided Student's *t*-test. **h**, $P < 0.0001$; **i**, $P = 0.0045$ and 0.0028 for laminae I–II and laminae III–V, respectively. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



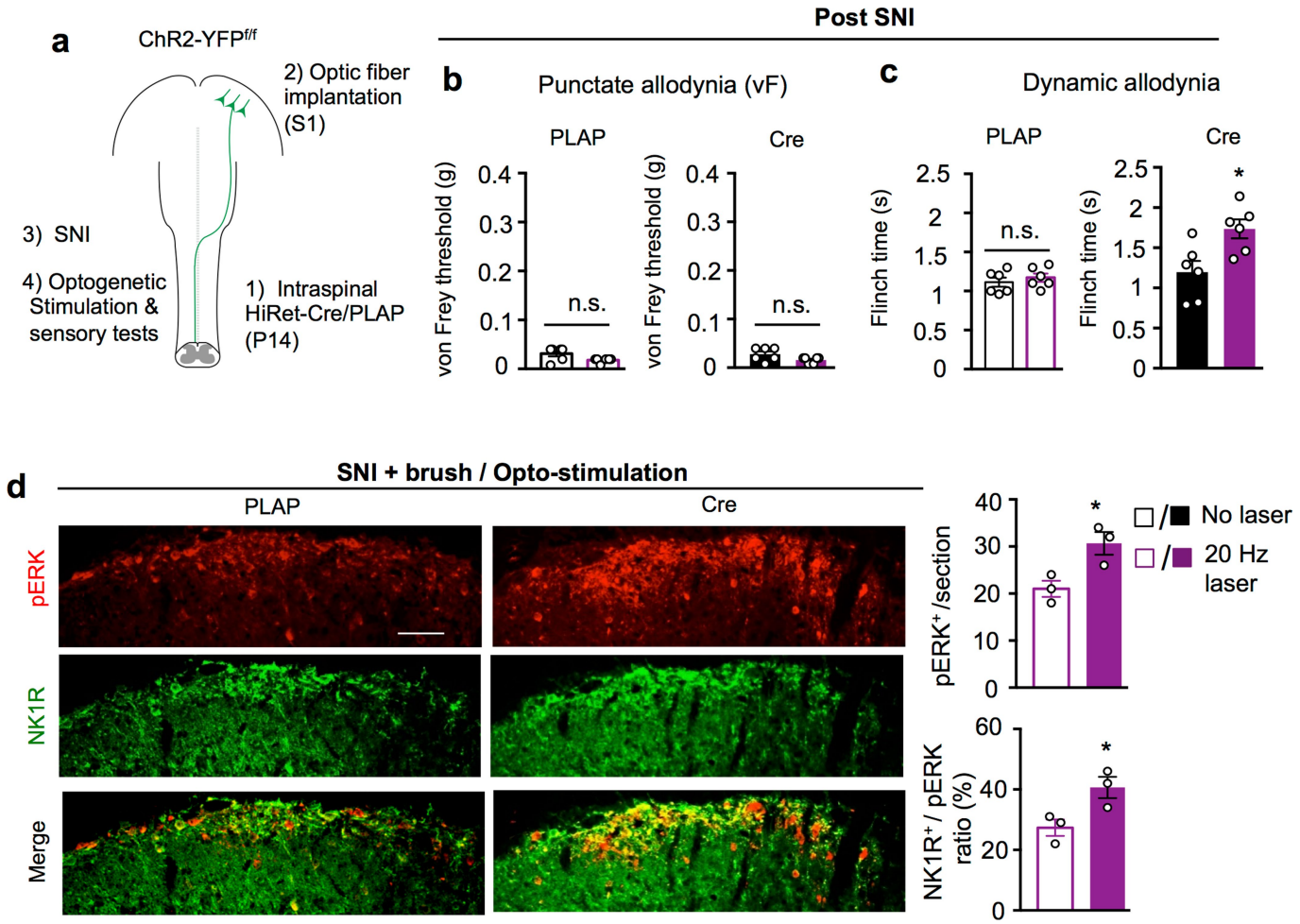
	eEPSC	APs	eIPSC	No response
CST	20/40	4/40	13/40	8/40
Aβ	20/40	4/40	17/40	11/40

Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | Characterization of Aβ fibre and CST inputs onto CCK–tdTomato neurons. a, Schematic of the stimulation and whole-cell patch recording set-up for tdTomato-labelled CCK^+ interneurons. CST axons labelled by AAV-ChR2-YFP were stimulated with a 473-nm laser. A single dorsal root (L4–L6) was stimulated with a glass suction electrode. **b**, **c**, Representative consecutive traces ($n = 3$) of Aβ fibre (left) and opto-CST (right) stimulation-evoked responses (**b**) and summarization (**c**) of whole cell patch-clamp recordings on CCK–tdTomato interneurons. Three recording conditions were used: first, to detect evoked EPSCs (eEPSCs), we held the membrane potential at −70 mV, which is the equilibrium potential of Cl^- and therefore minimizes the flow of eIPSCs. Second, by holding the membrane potential at 0 mV, we examined the polysynaptic, inhibitory inputs (eIPSCs) onto CCK–tdTomato interneurons. Third, we used current clamp mode to examine whether the stimulation drove action potential firing at the resting membrane potential. Type 1: CCK–tdTomato neurons receive only excitatory inputs, and few of them generated AP outputs when Aβ or CST inputs were stimulated. Type 2: CCK–tdTomato neurons receive both excitatory inputs and feed-forward inhibitory inputs, with no AP output. Type 3: CCK–tdTomato neurons receive predominant feed-forward inhibition, with no AP output. Type 4: CCK–tdTomato neurons show no response at either voltage or current clamp recording. **d**, Left, representative recording of an Aβ (25 μA) dorsal root evoked EPSC at −70 mV. Latency and jitter properties (magnified in inset) with quantifications ($n = 8$ neurons) are consistent with monosynaptic sensory connectivity. Right, opto-stimulation evoked EPSCs (averaged traces) at −70 mV in the same cell as shown on the left. The evoked EPSC was blocked by AMPA and NMDA antagonists (NBQX (5 μM)/CPP (20 μM)). In addition, opto-stimulation evoked EPSCs were eliminated by TTX (0.5 μM), and reinstated by 4-AP (2 mM), indicating a monosynaptic connection between the CST and CCK–tdTomato interneurons. Bar graph, quantification of eEPSC amplitudes after administration of drugs as shown. $^{**}P < 0.0001$; one-way ANOVA followed by Bonferroni correction. $n = 8$ neurons. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Reinforcement of tactile allodynia in mice with SNI by optogenetic stimulation of somatosensory CSNs. a, Drawing of experimental paradigm. **b**, **c**, Measurement of punctate (**b**, von Frey filaments) and dynamic (**c**, brush) mechanical allodynia upon opto-stimulation in control mice ($n = 6$) and mice expressing ChR2-YFP in CSNs ($n = 6$) after SNI. n.s., no statistical significance; $^*P < 0.05$. **b**, **c**, $P = 0.18$, 0.08 and 0.41, 0.02 for PLAP and Cre, without or with laser, respectively; two-sided Student's *t*-test. **d**, Representative images and quantification of pERK (red) and NK1R (green) immunostaining in the superficial dorsal horn (laminae I–II) of the spinal cord (L3–4) in control mice ($n = 3$) or mice expressing ChR2-YFP in hindlimb CSNs ($n = 3$) with SNI and brush stimulation coupled with opto-stimulation. Scale bar, 100 μm. $^*P < 0.05$; $P = 0.02$ and 0.01 for pERK and pERK/NK1R ratio, respectively; two-sided Student's *t*-test. Six sections crossing the lumbar spinal cord (L3–4) were quantified for individual animals. Data shown as mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0534-z

Nuclear envelope assembly defects link mitotic errors to chromothripsis

Shiwei Liu[1,2,3,5], Mijung Kwon[1,2,3,5], Mark Mannino[1,2,3], Nachen Yang[4], Fioranna Renda[4], Alexey Khodjakov[4] & David Pellman[1,2,3]*

Defects in the architecture or integrity of the nuclear envelope are associated with a variety of human diseases[1]. Micronuclei, one common nuclear aberration, are an origin for chromothripsis[2], a catastrophic mutational process that is commonly observed in cancer[3–5]. Chromothripsis occurs after micronuclei spontaneously lose nuclear envelope integrity, which generates chromosome fragmentation[6]. Disruption of the nuclear envelope exposes DNA to the cytoplasm and initiates innate immune proinflammatory signalling[7]. Despite its importance, the basis of the fragility of the micronucleus nuclear envelope is not known. Here we show that micronuclei undergo defective nuclear envelope assembly. Only 'core' nuclear envelope proteins[8,9] assemble efficiently on lagging chromosomes, whereas 'non-core' nuclear envelope proteins[8,9], including nuclear pore complexes (NPCs), do not. Consequently, micronuclei fail to properly import key proteins that are necessary for the integrity of the nuclear envelope and genome. We show that spindle microtubules block assembly of NPCs and other non-core nuclear envelope proteins on lagging chromosomes, causing an irreversible defect in nuclear envelope assembly. Accordingly, experimental manipulations that position missegregated chromosomes away from the spindle correct defective nuclear envelope assembly, prevent spontaneous nuclear envelope disruption, and suppress DNA damage in micronuclei. Thus, during mitotic exit in metazoan cells, chromosome segregation and nuclear envelope assembly are only loosely coordinated by the timing of mitotic spindle disassembly. The absence of precise checkpoint controls may explain why errors during mitotic exit are frequent and often trigger catastrophic genome rearrangements[4,5].

During telophase, the assembling nuclear envelope (NE) transiently forms two domains around decondensing chromosomes[1,8,9]. Core NE proteins (for example, the membrane protein emerin) assemble all around chromosomes and then concentrate in regions adjacent to the spindle. Concomitantly, NPCs and other non-core proteins (for example, the lamin B receptor LBR) assemble on the chromosome periphery away from the spindle and are transiently excluded from the core domain[8,9] (Extended Data Fig. 1a–c). After mitotic exit, these domains are intermingled and additional NPCs slowly assemble via an interphase assembly pathway[10].

To understand the NE defect of micronuclei (MN), we investigated whether NE assembly on lagging chromosomes follows the same spatiotemporal pattern as the primary nuclei (PN). We generated lagging chromosomes in synchronized cells of three cell lines by two independent methods (see Methods). Whereas core proteins were recruited to lagging chromosomes at equivalent or higher levels than to the main chromosome mass, NPCs and other non-core proteins almost completely failed to assemble (Fig. 1a, Extended Data Fig. 1a–f). Chromatin bridges extending between daughter cells displayed a similar defect in NE protein composition (Extended Data Fig. 1g, h). Thus, consistent with a recent study[11], only a subset of NE proteins assembles normally on chromatin that remains within the telophase central spindle.

MN from lagging chromosomes also showed reduced assembly of NPCs and non-core proteins (Extended Data Fig. 2). Accordingly, with some variability in penetrance, MN had substantial nuclear import defects (Fig. 1b, c, Extended Data Fig. 3a, b, Supplementary Videos 1, 2) and were defective in accumulating nuclear proteins such as replication protein A (RPA), a key dosage-sensitive regulator of DNA replication and repair[12], and B-type lamins, which are required to maintain NE integrity[6,13] (Extended Data Figs. 2, 3c–e). For individual cells, the amount of RPA and an importin beta-dependent reporter were correlated (Extended Data Fig. 3e), suggesting that there is a general import defect, at least for proteins imported by importin beta. Thus, reconciling differing prior reports[6,11,14–16], MN undergo aberrant NE and NPC assembly, leading to defective import and accumulation of nuclear proteins.

We considered the possibility that the defect in assembly of NPCs and non-core NE proteins on lagging chromosomes was due to a previously proposed 'chromosome separation checkpoint'[15]. Under this model, the spindle midzone-centred phosphorylation gradient that is generated by the mitotic kinase aurora B[17] provides a spatial cue that blocks NE assembly until membrane-free lagging chromosomes can be incorporated within the main chromosome mass to form a single nucleus. However, several observations were inconsistent with this model. First, core membrane proteins, which were not previously examined[15], assemble on lagging chromosomes and chromosome bridges independent of their position relative to the spindle midzone (Extended Data Fig. 4a), indicating that an NE does form on lagging chromosomes, albeit one lacking NPCs. Second, chromosome bridges are uniformly depleted of non-core NE proteins, displaying no obvious gradient (Extended Data Fig. 1g, h). Third, in late telophase, lagging chromosomes lose aurora B-mediated phosphorylation of histone 3 serine 10 (pH3S10) but do not assemble NPCs or other non-core proteins (Extended Data Fig. 4b, c). Finally, in cells where the midzone aurora B gradient was lost after short interfering RNA (siRNA)-mediated knockdown of its transport motor MKLP2[17], non-core NE assembly was still inhibited on lagging chromosomes (Extended Data Fig. 4d).

The checkpoint model requires that lagging chromosome position be monitored continuously throughout mitotic exit by aurora B. To test this idea, cells were released from mitotic arrest, imaged with high temporal resolution, treated with an aurora B inhibitor (ZM447439) for 12 min around anaphase or telophase, and then fixed and labelled to assess NE assembly (Extended Data Fig. 4e). Because cells release from the mitotic block with only partial synchrony, we could identify cells in which aurora B was inhibited at each stage between metaphase and telophase. Unexpectedly, inhibition of aurora B restored non-core protein assembly to lagging chromosomes only if it occurred early, at or shortly after anaphase onset, but not later, about 6–8 min after anaphase onset (Fig. 2a). The inability of late aurora B inhibition to restore non-core NE assembly cannot be explained by a partial effect of drug inhibition: near-complete loss of aurora B activation was observed within 3 min of inhibitor treatment and indistinguishable results were obtained using

[1]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [2]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [3]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [4]Wadsworth Center, New York State Department of Health, Albany, NY, USA. [5]These authors contributed equally: Shiwei Liu, Mijung Kwon. *e-mail: David_Pellman@dfci.harvard.edu

© 2018 Springer Nature Limited. All rights reserved.

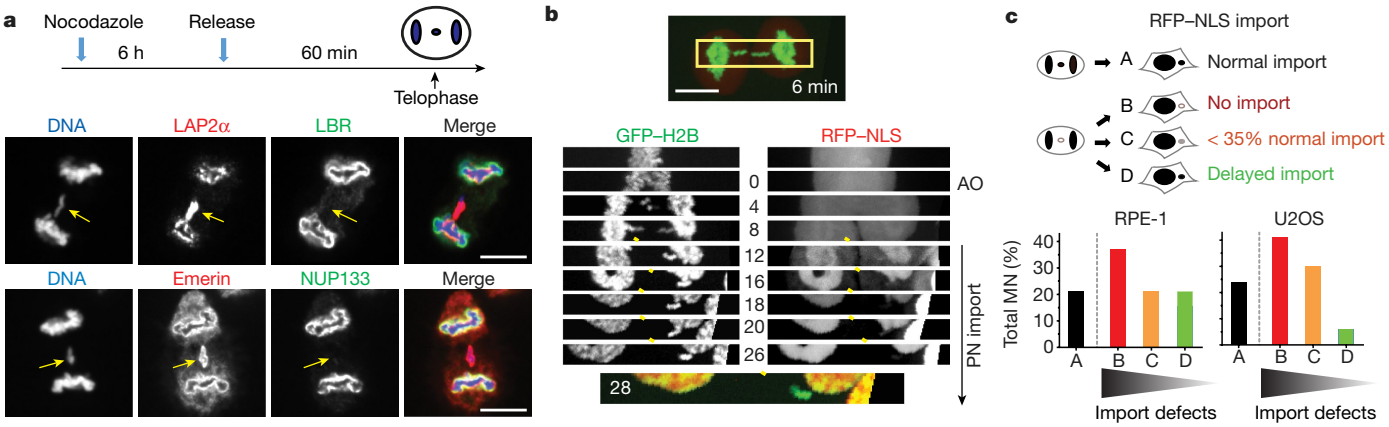


Fig. 1 | Defective NE assembly on lagging chromosomes. a, Non-core NE assembly defect on lagging chromosomes. Top, experimental scheme. Bottom, images of RPE-1 cells with lagging chromosomes (arrows, three experiments). Red, core NE proteins; green, non-core proteins. **b**, **c**, Impaired nuclear import in MN. **b**, Kymograph of an RPE-1 cell (boxed region of top image) shows impaired import of RFP fused to a nuclear localization signal (RFP–NLS) in the newly formed MN (arrowheads). Synchronization as in **a** ($t = 0$ is anaphase onset (AO)). Bottom, merged image. Representative images of nine cells, category B in **c**. **c**, Top, cartoon summarizing patterns of import to MN (see Extended Data Fig. 3a for representative quantification of import defects). Bottom, percentage of MN corresponding to the categories above ($n = 24$ from 11 experiments for RPE-1; $n = 17$ from 9 experiments for U2OS). Scale bars, 10 μm.

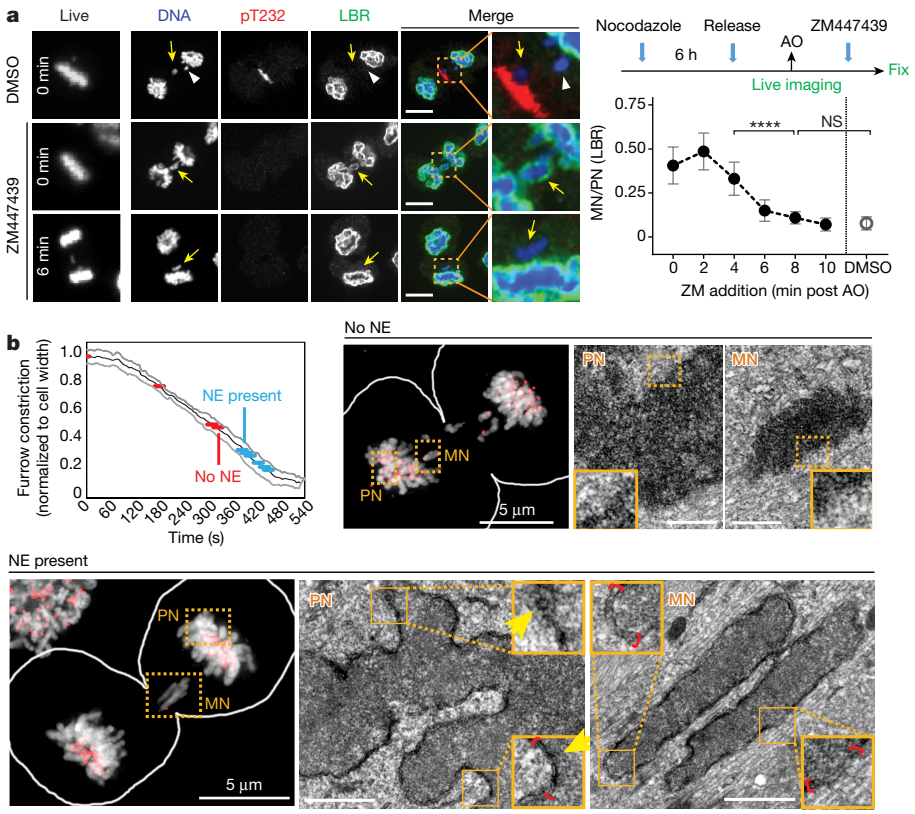


Fig. 2 | Irreversibility of the NE assembly defect on lagging chromosomes. a, Inhibition of aurora B after late anaphase fails to restore non-core (LBR) assembly to lagging chromosomes (arrows). Left, cells at the time of ZM447439 or DMSO addition (min after AO). Right columns, cells labelled for DNA (blue), phospho-T232 aurora B (red) and LBR (green). In controls, lagging chromosomes fail to recruit LBR even if they are located (white arrowhead) far from active aurora B (pT232). Scale bars, 10 μm. Right, top, experimental scheme. Right, bottom, MN/PN fluorescence intensity (FI) ratio for LBR in cells exposed to ZM447439 (ZM) at the indicated times (mean with 95% confidence interval (CI), $n = 40, 29, 31, 31, 43, 43, 34$ for data points from left to right, from five experiments. For brevity, lagging chromosomes are designated as MN). ****$P < 0.0001$; NS, $P = 0.0816$, two-tailed Mann–Whitney test. **b**, CLEM showing near-synchronous NE formation on all chromosomes. Top left, time course of furrow ingression in RPE-1 cells with lagging chromosomes (black line, mean; grey lines, ±1 s.d.). Red and blue discs indicate stages of the 10 cells analysed by CLEM. No NE is detected by CLEM during earlier stages of cytokinetic furrowing (red); at later stages (blue), double-membrane NE is present at both PN and MN. Light micrograph images: DNA in greyscale, kinetochores (CENPA–GFP) and centrosomes (centrin 1–GFP) in red. EM images: perinuclear space (red brackets in micrographs) is wider around MN than around PN. NPCs (arrowheads) are present only on PN. Scale bars, 0.5 μm.

© 2018 Springer Nature Limited. All rights reserved.

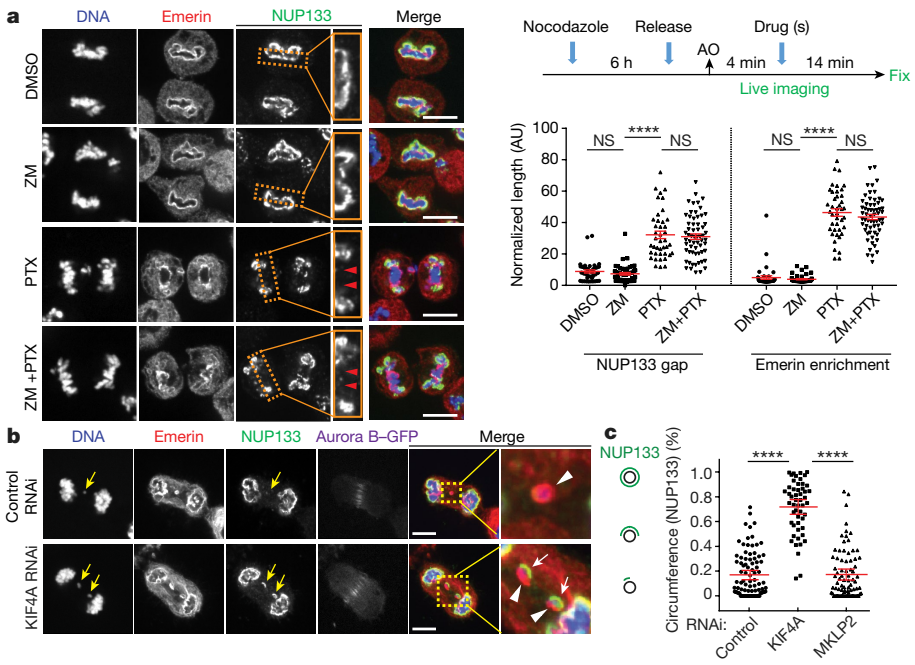


Fig. 3 | Spindle microtubules block non-core NE assembly independent of aurora B. **a**, Microtubule stabilization by paclitaxel (PTX) inhibits non-core invasion of core domain. Left, images of RPE-1 cells after drug treatment(s). Enlarged images from boxed regions show the non-core (NUP133) gap on the main chromosome mass (red arrowheads). Right, top, experimental scheme. Right, bottom, quantification of the results (mean with 95% CI, $n = 40, 47, 41, 61, 40, 47, 41, 61$, left to right, from two experiments). $^{****}P < 0.0001$; NS, $P = 0.2974, 0.9837, 0.2473, 0.3374$ (left to right); two-tailed Mann–Whitney test. **b**, **c**, Separation of small-scale core and non-core domains on lagging chromosomes (yellow arrows) after KIF4A depletion by RNA interference (RNAi). **b**, Images of HeLa K cells expressing aurora B–GFP (two technical replicates). Merged and enlarged image: emerin (red, white arrowheads), NUP133 (green, white arrows). Synchronization as in Extended Data Fig. 1e. **c**, The fraction of lagging chromosome circumference covered with NUP133 in HeLa K cells (left, cartoon; right, mean with 95% CI, $n = 90, 52, 81$, left to right, from two experiments). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. Scale bars, 10 μm.

a tenfold higher dose of the inhibitor (Extended Data Fig. 4f, g). Thus, aurora B inhibition has no effect after late anaphase, when a chromosome separation checkpoint would need to operate. Instead, lagging chromosomes become irreversibly defective for recruitment of non-core NE proteins around mid-anaphase.

Previous studies have suggested that the loss of chromosome condensation, marked by the loss of pH3S10 or condensins, might be required for NPC and NE assembly[15,18,19]. Accordingly, persistent mitotic condensin proteins have been suggested to block NPC and NE assembly on lagging chromosomes[15]. To further examine this, we used high-temporal-resolution imaging to relate the timing of NPC assembly to condensin levels. Unexpectedly, assembly of NPCs was readily observed with high levels of the condensin subunit SMC2 (about 70% of maximum; Extended Data Fig. 5a, b). Similar results were obtained for pH3S10, although there were antibody-specific differences in detecting the loss of pH3S10 (Extended Data Fig. 4c). In addition, about fivefold siRNA-mediated knockdown of SMC2 failed to restore NPC assembly on lagging chromosomes (Extended Data Fig. 5a, c). Moreover, in correlated light and electron microscopy (CLEM) experiments, all chromosomes, including lagging chromosomes, appeared comparably condensed when the NE assembled (Fig. 2b, bottom). Therefore, consistent with data from early cell fusion studies[20], extensive chromosome decondensation is not required for NPC assembly, and persistent condensin-binding cannot explain the irreversible NE assembly defect on lagging chromosomes.

Considering prior work[10,21,22], we hypothesized that the irreversible defect in NE assembly on lagging chromosomes might be caused by their enclosure in an NE that lacks NPCs. For example, *Xenopus* egg extracts lacking the NUP107–160 complex can assemble NPCs only if the purified complex is added back early in the reaction, because the formation of a continuous NPC-free NE irreversibly blocks NPC assembly[21]. Consistent with this idea, ESCRT-III components, which promote the closure of the newly formed NE in telophase[1,23], associated and dissociated from lagging chromosomes with normal kinetics (Extended Data Fig. 5d, e; Supplementary Video 3). Structured illumination microscopy (SIM) confirmed that lagging chromosomes were enclosed by the membrane protein emerin, but not by NPCs (Extended Data Fig. 5f). CLEM experiments demonstrated that double membranes formed near-synchronously around all chromosomes, including lagging chromosomes, although membranes around lagging chromosomes lacked NPCs (Fig. 2b). This paucity of NPCs persisted on interphase MN (Extended Data Fig. 5g), consistent with an irreversible defect in NE assembly.

Although our results argued against spatially patterned checkpoint regulation of NE assembly (above and Extended Data Figs. 4, 5), the effects of global aurora B inhibition on NE assembly in early anaphase still needed to be explained (Fig. 2a, Extended Data Fig. 4g). We considered the possibility that partial restoration of NE assembly might occur as an indirect consequence of aurora B-dependent stabilization of spindle microtubules during early anaphase[24] (Extended Data Fig. 6a). In *Xenopus* egg extracts, microtubule stabilization near chromosomes irreversibly inhibits the formation of an NE containing NPCs[22]. Accordingly, nocodazole and inhibition of aurora B had a similar effect on non-core NE assembly: microtubule disassembly by nocodazole during early anaphase reversed the non-core assembly defect on lagging chromosomes, whereas there was minimal effect if nocodazole was added later (Extended Data Fig. 6b, c). In addition, during normal NE assembly, non-core NE proteins are typically excluded from the core domain adjacent to dense microtubule bundles from the central spindle[8,25] (Extended Data Fig. 1b, c). However, this exclusion occurs irrespective of whether aurora B localizes to the spindle midzone or is forced to remain on the main chromosome mass (MKLP2 knockdown; Extended Data Fig. 6d).

Because of the interdependent functions of aurora B and microtubules[24], we used paclitaxel to prevent microtubule disassembly after inhibition of aurora B (Fig. 3a) and determined whether microtubules could still exclude the non-core NE from the central spindle region (Extended Data Figs. 1b, c, 6d). In control late-telophase cells, core and non-core proteins became intermingled on the main chromosome mass after mitotic exit, as expected. By contrast, in paclitaxel-treated cells, the gap between non-core NE domains persisted and/or the core domain became exaggerated (Fig. 3a, arrowheads). Notably, addition of an aurora B inhibitor together with paclitaxel had the same effect as paclitaxel treatment alone (Fig. 3a). Thus, microtubule inhibition of non-core NE protein assembly is independent of aurora B.

The above data suggest a model where non-core NE assembly is inhibited in areas with a high density of microtubules. To test this hypothesis, we 'loosened' spindle microtubule bundling by siRNA-mediated knockdown of the kinesin KIF4A[26]. As expected, KIF4A knockdown preserved general spindle organization and detectable central spindle localization of aurora B[26] (Extended Data Fig. 7a). However, KIF4A depletion partially reversed the defect in NPC and non-core NE assembly on lagging chromosomes (Fig. 3b, c, Extended Data Fig. 7b–d).

© 2018 Springer Nature Limited. All rights reserved.

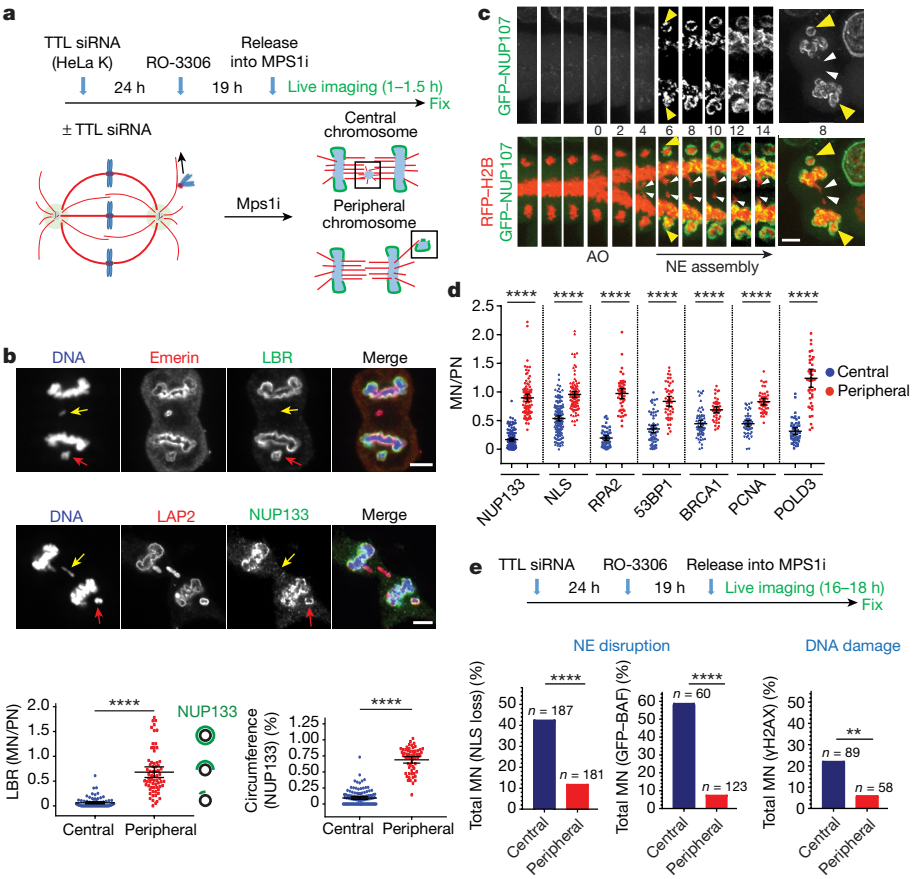


Fig. 4 | Peripheral localization of missegregated chromosomes corrects defects of MN. a, Experimental scheme to generate missegregated chromosomes positioned within or away from the spindle (microtubules, red; non-core NE, green). **b**, **c**, Non-core NE assembles on peripheral chromosomes but not central chromosomes. **b**, Top, images of RPE-1 cells with central (yellow arrows) and peripheral (red arrows) chromosomes labelled with core (red) or non-core (green) proteins. Bottom, quantification as in Fig. 3c (mean with 95% CI, $n = 83, 67$ (left), $n = 122, 58$ (right), from two experiments each). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. **c**, Similar result as in **b** from live-cell imaging of a HeLa K cell expressing RFP–H2B (red) and GFP–NUP107 (green) (Supplementary Video 4, representative of 15 cells). **d**, **e**, Functional restoration of MN from peripheral chromosomes. **d**, FI ratios for the indicated proteins, comparing MN from peripheral and central chromosomes (RPE-1 cells, scheme as in **a**). $^{****}P < 0.0001$, two-tailed Mann–Whitney test (mean with 95% CI, $n = 130, 97, 145, 112, 52, 47, 64, 48, 64, 47, 49, 45, 49, 45$, left to right, from two experiments for the indicated proteins). **e**, Top, experimental scheme. Bottom, graphs of the results (HeLa K cells). Left, NE integrity monitored by loss of NLS–RFP (five experiments). Middle, NE integrity monitored by hyperaccumulation of GFP–BAF (three experiments). Right, DNA damage (three experiments). $^{**}P = 0.0084$, $^{****}P < 0.0001$, two-tailed Fisher's exact test. Note, only a fraction of ruptured MN acquire DNA damage. Scale bars, 5 μm.

Notably, the NE at these sites often displayed small-scale separation of core (emerin, white arrowheads) and non-core proteins (NUP133, white arrows in Fig. 3b), as occurs on the main chromosome masses in telophase. These 'mini' core and non-core subdomains formed at any location within the central spindle, including the spindle midzone where aurora B activity is high. Moreover, these partially corrected lagging chromosomes exhibited high levels of pH3S10, which are comparable to those of metaphase chromosomes (Extended Data Fig. 7e). This further supports that NPC assembly can occur on condensed chromosomes. In addition, SIM after KIF4A knockdown suggested that the 'core-only' region of the lagging chromosome was enriched for NE–microtubule contacts marked by the ESCRT-III component CHMP4B[23] (Extended Data Fig. 8a–c). Notably, inhibition of aurora B in early anaphase has a very similar effect to KIF4A knockdown (Extended Data Fig. 8d), with mini core and non-core subdomains often forming around lagging chromosomes. Together, these data suggest that the local organization of microtubules is the key determinant of whether NPC and non-core NE proteins can assemble.

Our microtubule-inhibition model predicts that positioning missegregated chromosomes away from the central spindle should normalize NE assembly and restore function to MN. To generate peripheral missegregated chromosomes, we treated cells with an MPS1 kinase inhibitor, which caused missegregation before chromosome congression was complete. We combined MPS1 inhibition in HeLa cells with knockdown of tubulin tyrosine ligase (TTL), which further enhances peripheral chromosome localization[27] (Fig. 4a, Extended Data Fig. 9a). Strikingly, peripheral chromosomes recruited both core and non-core NE proteins, including NPCs, unlike central lagging chromosomes (Fig. 4a–c, Extended Data Fig. 9a, Supplementary Video 4). Labelling of microtubules and ESCRT-III[23] indicated that peripheral chromosomes also had less contact with microtubules than central chromosomes (Extended Data Fig. 9b). MN derived from peripheral chromosomes had normal levels of nuclear import and restored levels of lamin B1 and other nuclear proteins required for DNA replication and repair (Fig. 4d, Extended Data Fig. 9c, d). Moreover, MN derived from peripheral chromosomes exhibited a normal extent of DNA replication (Extended Data Fig. 9e). The restoration of a functional NE on peripheral MN, which occurred independent of their chromosome number, is likely to contribute to their larger size (Extended Data Fig. 9c–h). Finally, in HeLa cells, MN from peripheral chromosomes had an approximately fivefold lower rate of NE disruption and significantly less DNA damage ($P = 0.0084$) than MN from central chromosomes (Fig. 4e, Extended Data Fig. 10a, b). In RPE-1 cells, peripheral chromosome localization also significantly reduced NE disruption and DNA damage ($P < 0.0001$, Extended Data Fig. 10c, Supplementary Video 5). However, in RPE-1 cells, the effect of

© 2018 Springer Nature Limited. All rights reserved.

peripheral chromosomes on NE disruption was partly masked because the large MN from peripheral chromosomes in these cells are subject to actin-dependent NE breakage, as can occur transiently on PN[13] (Extended Data Fig. 10c, d). Thus, positioning chromosomes away from the spindle restores near-normal NE assembly and function of the resulting MN.

We have explained the NE defects of MN. Our findings define an important mechanism underlying chromothripsis and suggest a new model for the coordination of chromosome segregation and NE assembly during normal cell division (Supplementary Video 6). We propose that, during telophase, bundled spindle microtubules inhibit the recruitment of NE with NPCs and non-core proteins to lagging chromosomes. Lagging chromosomes are therefore enclosed by an NE primarily containing core proteins. Prior in vitro studies[21,22] suggest that this NPC-deficient NE irreversibly blocks the assembly of additional NPCs and non-core proteins. This irreversible defect is likely to be accentuated by sequestration of non-core proteins by the newly forming daughter PN. Because MN start out with an NE containing few NPCs, they have reduced access to interphase NPC assembly, which requires nuclear transport[10]. Defective nucleocytoplasmic transport then leads to defects in the accumulation of numerous proteins, including those necessary for normal DNA replication, DNA repair, and the maintenance of NE integrity (for example, B-type lamins[6,13]; Extended Data Figs. 9d, 10d). The resulting spontaneous NE disruption leads to chromosome fragmentation and ultimately chromothripsis[2,6,14]. How microtubules inhibit non-core NE assembly remains to be determined, but it could occur, at least in part, through a simple physical barrier effect. NPC precursors have recently been reported to assemble on fenestrated endoplasmic reticulum sheets during telophase[28] (and personal communication, T. Kirchhausen); we speculate that they might less readily penetrate bundles of spindle microtubules as a result.

Our findings demonstrate that altered NE assembly on lagging chromosomes is not the consequence of a beneficial checkpoint delay, but rather a pathological outcome. During normal cell division, instead of precise and continuous monitoring of chromosome position, it appears that there is only loose coordination, with normal non-core NE assembly being dependent on timely disassembly of spindle microtubules. Loose coordination between chromosome segregation and NE assembly, coupled with the irreversibility of NE assembly errors during mitotic exit, provides an explanation of why chromothripsis is common, with frequencies recently reported to be as high as 65% in some cancers[4,5].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0534-z.

Received: 30 January 2018; Accepted: 4 August 2018;
Published online 19 September 2018.

1. Ungricht, R. & Kutay, U. Mechanisms and functions of nuclear envelope remodelling. *Nat. Rev. Mol. Cell Biol.* **18**, 229–245 (2017).
2. Zhang, C.-Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
3. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
4. Notta, F. et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
5. Ly, P. & Cleveland, D. W. Rebuilding chromosomes after catastrophe: emerging mechanisms of chromothripsis. *Trends Cell Biol.* **27**, 917–930 (2017).
6. Hatch, E. M., Fischer, A. H., Deerinck, T. J. & Hetzer, M. W. Catastrophic nuclear envelope collapse in cancer cell micronuclei. *Cell* **154**, 47–60 (2013).
7. Umbreit, N. T. & Pellman, D. Cancer biology: Genome jail-break triggers lockdown. *Nature* **550**, 340–341 (2017).
8. Haraguchi, T. et al. Live cell imaging and electron microscopy reveal dynamic processes of BAF-directed nuclear envelope assembly. *J. Cell Sci.* **121**, 2540–2554 (2008).
9. Dechat, T. et al. LAP2α and BAF transiently localize to telomeres and specific regions on chromatin during nuclear assembly. *J. Cell Sci.* **117**, 6117–6128 (2004).
10. Otsuka, S. et al. Nuclear pore assembly proceeds by an inside-out extrusion of the nuclear envelope. *eLife* **5**, e19071 (2016).
11. de Castro, I. J., Gil, R. S., Ligammari, L., Di Giacinto, M. L. & Vagnarelli, P. CDK1 and PLK1 coordinate the disassembly and reassembly of the nuclear envelope in vertebrate mitosis. *Oncotarget* **9**, 7763–7773 (2017).
12. Toledo, L. I. et al. ATR prohibits replication catastrophe by preventing global exhaustion of RPA. *Cell* **155**, 1088–1103 (2013).
13. Hatch, E. M. & Hetzer, M. W. Nuclear envelope rupture is induced by actin-based nucleus confinement. *J. Cell Biol.* **215**, 27–36 (2016).
14. Crasta, K. et al. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**, 53–58 (2012).
15. Afonso, O. et al. Feedback control of chromosome separation by a midzone aurora B gradient. *Science* **345**, 332–336 (2014).
16. Karg, T., Warecki, B. & Sullivan, W. Aurora B-mediated localized delays in nuclear envelope formation facilitate inclusion of late-segregating chromosome fragments. *Mol. Biol. Cell* **26**, 2227–2241 (2015).
17. Fuller, B. G. et al. Midzone activation of aurora B in anaphase produces an intracellular phosphorylation gradient. *Nature* **453**, 1132–1136 (2008).
18. Ramadan, K. et al. Cdc48/p97 promotes reformation of the nucleus by extracting the kinase aurora B from chromatin. *Nature* **450**, 1258–1262 (2007).
19. Vagnarelli, P. et al. Repo-Man coordinates chromosomal reorganization with nuclear envelope reassembly during mitotic exit. *Dev. Cell* **21**, 328–342 (2011).
20. Obara, Y., Chai, L. S., Weinfeld, H. & Sandberg, A. A. Synchronization of events in fused interphase-metaphase binucleate cells: progression of the telophase-like nucleus. *J. Natl. Cancer Inst.* **53**, 247–259 (1974).
21. Walther, T. C. et al. The conserved Nup107-160 complex is critical for nuclear pore complex assembly. *Cell* **113**, 195–206 (2003).
22. Xue, J. Z., Woo, E. M., Postow, L., Chait, B. T. & Funabiki, H. Chromatin-bound *Xenopus* Dppa2 shapes the nucleus by locally inhibiting microtubule assembly. *Dev. Cell* **27**, 47–59 (2013).
23. Vietri, M. et al. Spastin and ESCRT-III coordinate mitotic spindle disassembly and nuclear envelope sealing. *Nature* **522**, 231–235 (2015).
24. Hochegger, H., Hégarat, N. & Pereira-Leal, J. B. Aurora at the pole and equator: overlapping functions of aurora kinases in the mitotic spindle. *Open Biol.* **3**, 120185 (2013).
25. Lu, L., Ladinsky, M. S. & Kirchhausen, T. Formation of the postmitotic nuclear envelope from extended ER cisternae precedes nuclear pore assembly. *J. Cell Biol.* **194**, 425–440 (2011).
26. Kurasawa, Y., Earnshaw, W. C., Mochizuki, Y., Dohmae, N. & Todokoro, K. Essential roles of KIF4 and its binding partner PRC1 in organized central spindle midzone formation. *EMBO J.* **23**, 3237–3248 (2004).
27. Barisic, M. et al. Mitosis. Microtubule detyrosination guides chromosomes during mitosis. *Science* **348**, 799–803 (2015).
28. Otsuka, S. et al. Postmitotic nuclear pore assembly proceeds by radial dilation of small membrane openings. *Nat. Struct. Mol. Biol.* **25**, 21–28 (2018).

Acknowledgements We thank I. Cheeseman, T. Rapoport, N. Umbreit, T. Walther, and K. Xie for advice or comments on the manuscript; E. Jackson and A. Spektor for preliminary experiments; R. Davidowitz for animations; J. Ellenberg, D. Gerlich, E. Hatch, M. Hetzer, A. Hyman, T. Kuroda, and L. Shao for reagents; and J. Waters and T. Lambert of the Nikon Imaging Center at Harvard Medical School and H. Elliott and C. Yapp from the Image and Data Analysis Core at Harvard Medical School for advice. We acknowledge the use of the Wadsworth Center's Electron Microscopy Core Facility. A. K. is supported by NIH GM059363. D.P. is a HHMI investigator and is supported by R37 GM61345.

Author contributions D.P., S.L., and M.K. designed the experiments. D.P., S.L., and M.K. wrote the manuscript, with edits from all authors. S.L. and M.K. performed most experiments and analysis. M.M. assisted with several experiments and contributed Extended Data Figs. 4g, 5b, 8d, 10c, d. N.Y. and A.K. performed the electron microscopy in Extended Data Fig. 5g. F.R. and A.K. performed the electron microscopy and contributed Fig. 2b.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0534-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0534-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Cell culture. Cell lines were maintained at 37 °C with 5% CO_2 atmosphere in Dulbecco's modified Eagle's medium (DMEM) (HeLa Kyoto (HeLa K), U2OS) or DMEM:F12 without phenol red (hTERT RPE-1). All media were supplemented with 10% FBS, 100 IU ml^{-1} penicillin, and 100 μg ml^{-1} streptomycin. Stable cells lines (H2B–eGFP and TDRFP–NLS RPE-1, mRFP–H2B and eGFP–BAF RPE-1, mRFP–H2B RPE-1, eGFP–BAF and mRFP–LAP2β RPE-1, mRFP–H2B and mNeonGreen–PCNA RPE-1, H2B–eGFP and TDRFP–NLS HeLa K, eGFP–BAF HeLa K, TDRFP–NLS U2OS) were generated by transduction using lentivirus or retrovirus vectors carrying the genes of interest. Cells were infected for 18–24 h with virus in the presence of 10 μg ml^{-1} polybrene, washed, and allowed to recover before antibiotic selection or fluorescence-activated cell sorting (FACS). Virus was generated by transfection of HEK293FT cells with the appropriate packaging plasmids (lentivirus, pMD2.G and psPAX2; retrovirus, pUMVC and pVSV-G) using Lipofectamine 3000 (Life Technologies) according to the manufacturer's instructions. RPE-1 cells stably expressing emerin–eGFP were generated by transient transfection (Lipofectamine 3000 transfection reagent, Life Technologies) of emerin–eGFP plasmid followed by selection using G418 (500 μg ml^{-1}). HeLa K cells stably expressing mRFP–H2B and CHMP4B–eGFP (BAC), H2B–mCherry and aurora B–eGFP (BAC) were derived from CHMP4B–eGFP HeLa K cells and aurora B–eGFP HeLa K cells, respectively (gifts from A. Hyman)[29]. HeLa K cells stably expressing mRFP–H2B and mEGFP–NUP107 were derived from mEGFP–NUP107-expressing HeLa K cells (constructed by zinc finger nuclease-mediated genome-editing, from J. Ellenberg)[10]. U2OS cells stably expressing H2B–mCherry and 3×GFP–IBB were a gift from E. Hatch and M. Hetzer[6]. All cells and the derivative cell lines for this study were tested for mycoplasma.

DNA constructs. The following constructs were obtained from Addgene: H2B-eGFP (plasmid 11680, from G. Wahl), emerin-eGFP (plasmid 61985, from E. Schirmer), pmRFP-LAP2β-IRES-puro2b (plasmid 21047, from D. Gerlich). The construct encoding eGFP–BAF was made by recombining full-length human BAF into a pLenti-CMV-neo-eGFP vector (T. Kuroda, vector plasmid Addgene 17447, from E. Campeau and P. Kaufman). The construct encoding mRFP–H2B was generated by recombining mRFP–H2B into a pBABE-Puro vector (T. Kuroda, vector plasmid Addgene 1764, from H. Land, J. Morgenstern and B. Weinberg). The construct encoding TDRFP–NLS (referred to as RFP–NLS throughout) was a gift from A. Salic. The construct encoding mNeonGreen–PCNA was made by recombining mNeonGreen–PCNA into a pLENTI-CMV-Neo-DEST vector (N. Umbreit, vector plasmid Addgene 17392, from E. Campeau and P. Kaufman) using Gateway LR Clonase II Enzyme (Invitrogen). Before Gateway cloning (Invitrogen), mNeonGreen–PCNA was first amplified by PCR using Ex Taq polymerase (Takara, primers: forward 5′ CACCATGGTGAGCAAGGG 3′; reverse 5′ CCTCTACAAATGTGGTATGGCTG 3′) and then the resulting PCR product was inserted into the pCR8/GW/TOPO vector (Invitrogen), according to the manufacturer's instructions.

Antibodies for immunofluorescence. For immunofluorescence, the following primary antibodies were used: LAP2α (1:300, Abcam, ab66588), LAP2β (1:200, Bethyl Laboratories, A304-840A), LAP2 (1:750, BD Transduction, 611000), mouse emerin (1:300, Abcam, ab54996), rabbit emerin (1:150, Proteintech, 10351-1-AP), rabbit lamin A/C (1:1,000, Abcam, ab26300), mouse lamin A (1:1,000, Abcam, ab8980), rabbit LBR (1:500, Abcam, ab32535), mouse LBR (1:300, Sigma-Aldrich, SAB1400151), NUP133 (1:300, Abcam, ab155990), NUP107 (1:100, Life Technologies, 39C7), NUP62 (1:500, BD Transduction, 610497), NUP358 (1:1,000, Abcam, ab64276), TPR (1:200, Abcam, ab84516), NUP153 (1:500, Abcam, ab24700), mAb414 (1:1,000–1:2,000, Abcam, ab24609), ELYS/MEL28 (1:200, Bethyl Laboratories, A300-166A), lamin B1 (1:1,000, Abcam, ab16048), lamin B2 (1:1,000, Abcam, ab151735), RPA1 (1:100, Cell Signaling, 2267), RPA2 (1:100, Abcam, ab2175), LSD1 (1:200, Cell Signaling, 2184), Rb (1:400, Cell Signaling, 9309), aurora B (1:100, Abcam, ab3609), mouse phospho-H3S10 (1:10,000, Abcam, ab14955), rabbit phospho-H3S10 (1:5,000, Abcam, ab5176), SMC2 (1:1,000, Abcam, ab10412), aurora B phospho-T232 (1:500, Rockland Antibodies, 600-401-677), CHMP4B (1:100, Proteintech, 13683-1-AP), CHMP2A (1:100, Proteintech, 10477-1-AP), IST1 (1:100, Proteintech, 51002-1-AP), rabbit α-tubulin (1:1,000, Abcam, ab18251), mouse α-tubulin (1:300, Sigma-Aldrich, DM1A), CENP-A (1:300, Abcam, ab13939), rabbit γH2AX (1:500, Cell Signaling, 2577), mouse γH2AX (1:500, EMD Milipore, JBW301), FluoTag (Atto488)-X4 anti-GFP (1:250, NanoTag Biotechnologies), 53BP1 (1:100, Cell Signaling, 4937), BRCA1 (1:200, Santa Cruz, D-9) and POLD3 (1:100, Abnova, Clone 3-E2). Secondary antibodies used were Alexa Fluro 405, 488, 568 and 647 (1:1,000, 1:500 for Alexa Fluro 405; Life Technologies).

SDS–PAGE and western blotting. Cells were lysed in 2× sample buffer (100 mM TrisHCl pH 6.8, 4% SDS, 10% β-mercaptoethanol). The whole-cell lysate was subjected to SDS–polyacrylamide gel electrophoresis on NuPAGE 4–12% Bis-Tris gradient gels (Novex Life Technologies) and proteins were transferred onto a nitrocellulose membrane (Millipore). For immunoblotting, the following primary antibodies were used: rabbit anti-MKLP2 (1:1,000, Bethyl Laboratories, A300-879A), mouse anti-SMC2 (1:1,000, Abcam, ab10412), rabbit anti-KIF4A (1:500, Bethyl Laboratories, A301-074A), rabbit anti-TTL (1:2,000, Proteintech, 13618-1-AP), and mouse anti-α-tubulin (1:15,000, Sigma-Aldrich, DM1A).

RNA interference. Cells at 50% confluency were transfected with 5 nmol (Extended Data Fig. 1h) or 40 nmol siRNA for 6 h using the Lipofectamine 3000 transfection reagent (Life Technologies) according to the manufacturer's instructions. Transfections were performed 48 h before the induction of lagging chromosomes. The following siRNA sequences were used: *MKLP2*, 5′-AACC ACCUAUGUAAUCUCAUGUU-3′[30]; *KIF4A*, 5′-AAGCAGAUUG AAAGCCUAGAGUU-3′[26]; *TTL*, 5′-CAGCCACCAAUCAGUAACUUU-3′[27]; *SMC2*, SMARTpool: ON-TARGETplus SMC2 siRNA L-006836-01-0005 (Dharmacon); *TP53*, SMARTpool: ON-TARGETplus TP53 siRNA L-003329-00-0050, (J-003329-14) GAAAUUUGCGUGUGGAGUA, (J-003329-15) GUGCAGCUGUGGGUUGAUU, (J-003329-16) GCAGUCAGAUCCUAGCGUC, (J-003329-17) GGAGAAUAUUUCACCCUUC (Dharmacon). Non-targeting Silencer Select Negative Control No. 1 siRNA (Thermo Fisher Scientific, 4390844) or Luciferase, 5′-CGUACGCGGAAUACUUCGATT-3′ were used as controls.

EdU incorporation. Synchronized cells were continuously incubated with 10 μM EdU starting 1 h after mitotic exit. Detection of EdU incorporation was performed using Click-iT Plus EdU Alexa Fluor imaging kits 647 or 594 (Life Technologies) according to the manufacturer's instructions.

SiR-DNA labelling for live-cell imaging. Synchronized cells were treated with 1 μM of the cell-permeable DNA dye SiR-DNA (Cytoskeleton Inc.), and 1 μM of verapamil, a broad-spectrum efflux pump inhibitor, approximately 30 min–1 h before imaging.

Cell cycle synchronization and generation of lagging chromosomes/MN. For the nocodazole block and release protocol to induce lagging chromosomes and MN, cells were treated with 100 ng ml^{-1} nocodazole (Sigma-Aldrich) for 6 h. Arrested cells were then collected by a shake-off procedure and washed three times with PBS before replating. For experiments where cells were imaged within the first 3 h after the nocodazole block and release, mitotic cells were plated onto coverslips or imaging dishes that were pre-coated with 0.1 mg ml^{-1} poly-D-lysine hydrobromide (Sigma-Aldrich) to facilitate attachment of mitotic cells. To obtain anaphase or telophase cells, cells were fixed about 40–50 min or 90 min after nocodazole release for RPE-1 and HeLa K cells, respectively. For experiments in which micronucleated cells were imaged in mid–late interphase, cells after shake-off were plated onto uncoated coverslips.

For the induction of lagging chromosomes and MN by MPS1 inhibition, cells were plated on coverslips or imaging dishes for 16–24 h and then treated with 9 μM RO-3306 (RO; EMD Millipore) for another 19 h to arrest cells in the G2 phase. Arrested cells were then washed 7 times with pre-warmed medium containing 10% FBS and treated with the MPS1 inhibitor NMS-P715 (1 μM, EMD Milipore). To induce a high frequency of peripheral missegregated chromosomes and their resulting MN, RPE-1 cells were treated with 3 μM NMS-P715. For HeLa K cells, TTL knockdown was done with 40 nmol *TTL* siRNA, added 24 h before the RO-3306 treatment. These G2-arrested cells (19 h) were washed and released into 3 μM NMS-P715. We note that peripheral missegregated chromosomes can also be generated at a high frequency using the combination of 0.05 μM CENP-E inhibitor (GSK923295, Selleckchem) and 1 μM NMS-P715 (EMD Milipore; not shown[31]).

Indirect immunofluorescence microscopy. Cells were washed in PBS and fixed with 4% paraformaldehyde for 15 min (unless otherwise stated); cells were then extracted in PBS-0.5% Triton X-100 for 5 min, washed 3 times with PBS, blocked for 30 min in PBS containing 3% BSA (PBS-BSA) and incubated with primary antibodies diluted in PBS-3% BSA for 60 min. These samples were then washed 3 times for 5 min with PBS-0.05% Triton X-100. Primary antibodies were detected using species-specific fluorescent secondary antibodies (Life Technologies). Finally, the samples were washed 3 times for 5 min with PBS-0.05% Triton X-100 before addition of 2.5 μg ml^{-1} Hoechst 33342 to detect DNA. Prolong Diamond Antifade or SlowFade Diamond Antifade (Life Technologies) was used for mounting of all immunofluorescence samples.

For labelling of SMC2 (Extended Data Fig. 5a–c), cells were pre-extracted with 0.5% Triton X-100 in CSK buffer (1 M PIPES, 10 mM NaCl, 300 mM EGTA, 1 mM $MgCl_2$) for 5 min on ice, fixed with 4% paraformaldehyde for 15 min and then processed as described above.

For labelling of CHMP4B, CHMP2A, and IST1 (Extended Data Fig. 5e), cells were fixed with methanol for 5–10 min at –20 °C and then processed as described above.

For co-staining of α-tubulin and CHMP4B (Extended Data Fig. 9b), microtubule integrity was preserved by fixation of cells with 0.5% glutaraldehyde in MTSB (80 mM K-PIPES, 5 mM EGTA, 1 mM $MgCl_2$ (pH 6.8)) for 10 min followed by addition of 0.1% $NaBH_4$ for 10 min to quench unreacted glutaraldehyde. Cells were then washed three times with PBS, and processed as above.

© 2018 Springer Nature Limited. All rights reserved.

Images were acquired on a Nikon Ti-E inverted microscope (Nikon Instruments, Melville, NY) with a Yokogawa CSU-22 spinning disk confocal head with the Borealis modification. Laser excitation used 405-nm, 488-nm, 561-nm and 642-nm lasers. Images were acquired using ×60 or ×100 Plan Apo NA 1.4 oil objectives with a CoolSnapHQ2 CCD camera (Photometrics). Acquisition parameters, shutters, filter positions and focus were controlled using Metamorph software (Molecular Devices).

All immunofluorescence images shown in the manuscript are from a single focal plane, with the exception of Extended Data Figs. 4b, 5f, 7e, 9b, 9g, which are maximum intensity projections from a *z*-focal plane series.

Correlative live-cell/fixed cell imaging experiments. For the nocodazole block and release protocol to induce lagging chromosomes and MN, mitotic cells were collected and processed as above, but were plated on glass-bottomed gridded 35 mm imaging μ-dishes (ibidi) precoated with 0.1 mg ml^{-1} poly-D-lysine hydrobromide (Sigma-Aldrich). For the induction of lagging chromosomes and MN by MPS1 inhibition, cells were plated on ibidiTreat bottomed gridded 35 mm μ-dishes (ibidi) 6–24 h before RO-3306 treatment. Imaging dishes were mounted on a TE2000-E2 inverted microscope equipped with a Nikon Perfect Focus system. An Okolab cage incubator was used to maintain samples at 37 °C and 5% humidified CO_2. Fluorescence and DIC images of cells expressing mRFP–H2B, H2B–eGFP and RFP–NLS, H2B–eGFP and eGFP–BAF were acquired with a ×20 0.5 NA Plan Fluor objective. After live-cell imaging, cells were washed with PBS, fixed and processed for immunofluorescence on the same imaging dishes. Cells of interest were identified from the recorded movies using the H2B channel (unless otherwise stated), and their positions were determined based on the grid pattern on the dishes using the DIC images (see Extended Data Fig. 4e). Immunostained samples were then imaged by spinning disk confocal microscopy as described above.

To characterize the recruitment pattern of emerin and NUP133 onto lagging chromosomes (Extended Data Fig. 1b), or the loss of H3S10 phosphorylation or SMC2 on lagging chromosomes or interphase MN (Extended Data Figs. 4b, c, 5b), we imaged synchronized mRFP–H2B-expressing RPE-1 cells at 2-min intervals and then subjected them to fixation and labelling as described above.

For ZM447439 or nocodazole treatment experiments to test for the reversal of the non-core NE assembly defect on lagging chromosomes in telophase (Fig. 2a, Extended Data Figs. 4e, g, 6b, c, 8d), we imaged synchronized mRFP–H2B-expressing RPE-1 cells at 2-min intervals during mitosis and added the drugs or vehicle controls during live-cell imaging. For paclitaxel/ ZM447439 experiments to analyse non-core NE exclusion from the main chromosome mass (Fig. 3a), we imaged synchronized mRFP–H2B-expressing RPE-1 cells and treated them with drugs or vehicle controls as described above. Following 12–14 min of live-cell imaging, cells were processed for indirect immunofluorescence as described above. Drug concentrations were: ZM447439 (5 μM or 50 μM, Tocris), nocodazole (10 μM, Sigma-Aldrich), paclitaxel (10 μM, Life Technologies).

For long-term correlated live-cell/fixed cell imaging experiments to assay for interphase micronuclear defects (Fig. 4, Extended Data Figs. 3d, e, 9, 10), RPE-1 or HeLa K cells expressing the tagged fluorescent proteins of interest were first imaged at higher temporal resolution to monitor the location of missegregated chromosomes during mitotic exit (3–5-min intervals for 1–1.5 h after anaphase onset). All analyses were done only from chromosomes that remained consistently peripheral or central relative to the mitotic spindle until mitotic exit. After mitotic exit, MPS1 inhibitor (NMS-P715) was reduced to 0.5–1 μM, and cells were then imaged at 20–30-min intervals for 16–18 h. When SiR-DNA was used to label DNA for long-term imaging, SiR-DNA concentration was reduced to 0.25 μM after initial 1.5 h of imaging. For latrunculin A treatment (Extended Data Fig. 10c, d), latrunculin A (150 nM, Life Technologies) was added to RPE-1 cells approximately 1 h after mitotic exit and maintained for the 16–18-h period of live-cell imaging. After live-cell imaging, samples were processed for immunofluorescence as described above to assay for NE rupture, DNA damage or DNA replication. Alternatively, to assess the accumulation of NE proteins, nuclear proteins or the import reporter into early interphase MN (Fig. 4d, Extended Data Fig. 9), cells were imaged at 20–30-min intervals for 1–3 h after mitotic exit and then prepared for indirect immunofluorescence.

Image analysis for indirect immunofluorescence. Image analysis was done with ImageJ/FIJI software. Regions of interest (ROIs) were generated for lagging chromosomes or MN and the corresponding PN using customized ImageJ/FIJI macros. First, nuclear segmentation was performed using Li or Otsu thresholding from the DNA (Hoechst) image or the mRFP–H2B image. Second, the ImageJ/FIJI functions 'Watershed' and 'Erode' were used to the refine the nuclear segmentation. Third, the segmented regions containing nuclei of interest (MN and corresponding PN) were manually selected as ROIs using the ImageJ/FI function 'Wand Tool'. Based on these ROIs, the mean fluorescence intensities (FI) of labelled proteins were quantified from other channels. For quantification of nuclear proteins (RFP–NLS, RPA1, RPA2, LSD1, Rb, 53BP1, BRCA1, POLD3, mNeonGreen–PCNA), chromatin-associated proteins (LAP2α, eGFP–BAF, γH2AX), or labelled DNA (as well as EdU incorporation), ROIs were generated based on the segmented nuclear area from the best focal plane for the MN and PN. For quantification of NE proteins or NPCs, band-shaped ROIs around MN and PN were generated using ImageJ/FJI function 'Make band' after nuclear segmentation followed by the procedures described above.

To characterize the assembly of NE or NPC proteins on lagging chromosomes (Fig. 1a, Extended Data Fig. 1), lagging chromosomes were scored as being positive if the mean FI (after background subtraction) of the labelled proteins on lagging chromosomes was >3 s.d. above the mean cytoplasmic background. Only telophase cells that had near-complete assembly of non-core NE proteins on the main chromosome mass were scored.

To obtain the FI ratios of proteins in MN relative to PN (MN/PN), the mean FI were quantified from the defined ROIs. Background-subtracted MN/PN FI ratios were then calculated and are shown in corresponding graphs.

To characterize the levels of phospho-H3S10 and SMC2 on the main and lagging chromosomes (Extended Data Figs. 4b, c, 5b, 7e), images were acquired from a *z*-stack (~12 μm) with a 0.2 μm (or 0.5 μm) step size to cover the region of the cell that contained the chromosomes. Because both anti-phospho-H3S10 antibodies showed stronger recognition for the antigens on the surface of the chromosome mass than those within the chromosome mass, the maximum intensity projections of *z*-focal plane series were used to estimate the average FI of phospho-H3S10. In parallel, the average FI of phospho-H3S10 was directly quantified from lines manually drawn along the periphery of the main chromosome mass (PN) and lagging chromosomes (MN), yielding similar results (not shown). Note that measurements for FI of NUP133 in Extended Data Figs. 4c, 5b were taken from the same samples.

To calculate the fraction of the lagging chromosome circumference labelled for NUP133 (Figs. 3c, 4b, Extended Data Figs. 6c, 7d, 8d, 9a, h), we measured the length of a manually drawn line along the lagging chromosome periphery that was positive for NUP133 (FI >3 s.d. above the mean cytoplasmic background). The ratio of this value was then made with the circumference of the lagging chromosome determined from the DNA label. This ratio was used in circumstances (for example, KIF4A knockdown) where some late mitotic NPC (non-core) assembly was restored on lagging chromosomes, but was restricted to a local 'mini-domain' (for example, Fig. 3c, Extended Data Figs. 7d, 8d). Only lagging chromosomes with a near-continuous emerin signal were analysed. In parallel, the MN/PN FI ratios of NUP133 (for example, Extended Data Fig. 4g where measurements were taken from the same samples as in Extended Data Fig. 8d) were quantified to verify late mitotic NPC assembly from corresponding experiments, yielding similar results.

To analyse the effect of aurora B inhibition (ZM447439) on spindle microtubule density (Extended Data Fig. 6a), the FI of α-tubulin was measured along the line perpendicular to the long axis of the mitotic spindle.

For paclitaxel/ ZM447439 experiments to analyse non-core NE exclusion on the main chromosome mass (Fig. 3a), the NUP133 gap was measured as the cumulative length of individual small discontinuities in the NUP133 signal (FI <3 s.d. above the mean cytoplasmic background) along the main chromosome mass. A threshold was set at ≥5 pixels as a definition for the lower bound for NUP133 discontinuities. The length of emerin enrichment was measured as the cumulative length of lines manually drawn along the main chromosome mass where emerin was enriched (>twofold emerin FI on the chromosome periphery). Only cells that had a near-continuous emerin signal on the main chromosome mass were quantified. The length of the NUP133 gaps and the emerin enrichment were normalized to the DAPI signal circumference of the main chromosome mass, set to 100.

To analyse the level of phospho-H3S10 on lagging chromsomes after KIF4A knockdown (Extended Data Fig. 7e), linescan profiles of the FI of phospho-H3S10 and NUP107 were measured along the white dashed line.

For the experiments to measure the restoration of DNA replication in peripheral chromosome-derived MN (Extended Data Fig. 9e), the MN/PN FI ratios of EdU were measured as described above, and then normalized to the DNA label FI ratio. To measure the frequency of micronuclear NE disruption (Fig. 4e, Extended Data Fig. 10a–c), MN were scored as being disrupted if the FI ratio of RFP–NLS in MN relative to the cytoplasm dropped to ~1 or if the FI ratio of eGFP–BAF in MN relative to PN rose to >3. For DNA damage, MN were scored as being positive if the mean FI of γH2AX in MN was >3 s.d. above the mean FI of γH2AX in the PN. This threshold was also used to exclude rare cells in which the PN had substantial DNA damage (for example, as can occur in a fraction of cells after high dose inhibition of MPS1).

Confocal live-cell microscopy. In all confocal live-cell experiments, cells were synchronized to induce lagging chromosomes or MN using the nocodazole block and release protocol, as described above. After mitotic shake-off, cells were replated onto glass-bottomed 35-mm imaging μ-dishes (ibidi) precoated with 0.1 mg ml^{-1} poly-D-lysine hydrobromide (Sigma-Aldrich). All images were collected on a Nikon Ti inverted microscope equipped with a Yokogawa CSU-X1 spinning disk confocal head, Spectral Applied Precision LMM-5 with AOTF, a Hamamatsu ORCA ER cooled CCD camera and the Nikon Perfect Focus System.

© 2018 Springer Nature Limited. All rights reserved.

GFP was imaged using a 488-nm laser for excitation and a 525/50-nm emission filter; RFP was imaged using a 561-nm laser for excitation and a 620/60-emission filter. An Okolab cage incubator was used to maintain samples at 37 °C and 5% humidified CO_2. Image acquisition was controlled with MetaMorph software (Molecular Devices).

For live-cell imaging to measure the kinetics of nuclear import into MN (Fig. 1b, c, Extended Data Fig. 3a, b), RPE-1 cells expressing H2B–eGFP and RFP–NLS, or U2OS cells expressing GFP–IBB and mCherry–H2B, or U2OS cells expressing RFP–NLS labelled with SiR-DNA were prepared for imaging as described above. To consistently track lagging chromosomes and MN that moved in and out of focal plane over time, images were acquired from 7 *z*-sections with a 2-μm step size for RPE-1 or 12 *z*-sections with 1-μm step size for U2OS using a ×60 Plan Apo NA 1.4 objective lens at 1-min intervals during the 30 min–3 h window after nocodazole release.

For live imaging of ESCRT-III dynamics (Extended Data Fig. 5d) or NUP107 dynamics (Fig. 4c, Extended Data Fig. 7c), HeLa K cells expressing CHMP4B–eGFP and mRFP–H2B or HeLa K cells expressing GFP–NUP107 and mRFP–H2B were prepared as described above. Images were collected from 10 *z*-sections with a 2-μm step size at 1-min intervals using a ×60 Plan Apo NA 1.4 objective lens during the 1–3-h window after nocodazole release.

Nuclear import assay. To quantify nuclear import (Extended Data Fig. 3a), customized ImageJ/FIJI macros were used. ROIs for lagging chromosomes or MN and PN were generated as described in 'Image analysis for indirect immunofluorescence'. Based on these ROIs, the mean FI of RFP–NLS for PN/MN pairs along with non-cell background was quantified through *z*-stacks from time-lapse image series. The background-subtracted mean FI of RFP–NLS from the best focal plane for pairs of lagging chromosomes or MN and PN was determined (best focal plane was by maximum FI, which could be different for the MN and PN). This procedure was used to enable accurate measurement of the FI of MN that moved above and below PN during the course of imaging. Detection accuracy was verified manually. The measured values were then corrected for photobleaching, using the bleaching coefficient described below.

We used G2 cells to calculate the rate of photobleaching in both intact MN and PN under what we presume were steady-state conditions. The measured bleaching coefficients were then used to correct for bleaching during the nuclear import assay. Live-cell imaging of RPE-1 cells expressing H2B-eGFP and RFP–NLS was performed in cells 16 h after release from the nocodazale block, using the same imaging settings and conditions as were used for nuclear import assay. We measured the RFP–NLS FI in the cytoplasm, PN and MN, subtracting background outside the cell. To control for subtle cell-to-cell variability in RFP–NLS expression (post FACS), the FI in the PN was set to 100 in each cell. The normalized mean FI of RFP–NLS from 12 MN/PN pairs was then plotted over time. For each PN/MN pair, the plotted data were fit to an exponential curve with an offset: $y = a \times \exp(-bx) + c$, where b is the bleaching coefficient, a is the y value at t_0 before bleaching, and c is the y value at plateau after bleaching (camera noise). From this analysis, there was no statistically significant difference between the average bleaching coefficients for the PN and MN ($b = 0.05$ (both PN and MN), $P = 0.62$ (NS), Wilcoxon matched-pairs test). This bleaching correction method was validated by comparing the experimentally measured (RFP–NLS FI) values and the predicted bleaching corrected values over time using the above equation. The ratios of experimentally measured values relative to the predicted bleaching-corrected values remained constant (nearly 1).

An orthogonal wide-field imaging approach, enabling a larger sample size, confirmed defective nuclear accumulation of RFP–NLS and RPA2 in MN (Extended Data Fig. 3d, e). For these experiments, imaging was performed at 5-min intervals using a ×20 objective, followed by fixation and quantitative immunofluorescence as described above.

Correlative light and electron microscopy (CLEM). For CLEM analyses of telophase cells (Fig. 2b), RPE-1 cells expressing CENPA–GFP and centrin 1–GFP were treated with 3 μM nocodazole for 4 h, released into fresh medium, and mounted in Rose chambers. Live-cell imaging (17 cells) was performed at 5-s intervals from mid-late anaphase through completion of cytokinesis. Dynamics of furrow ingression were determined by plotting the extent of constriction (normalized to the cell width) over time (Fig. 2b, top left). This plot was used to identify the stage of cells selected for electron microscopy. Cells were fixed with 2.5% glutaraldehyde in PBS, pH7.4 for 30 min. Chromosomes were stained with Hoechst 33342 at 1 μg ml^{-1}. Multimode (DIC and fluorescence) light microscope data sets were obtained on a Nikon TE2000 microscope equipped with a ×100 1.45 NA Plan Apo objective lens and Zyla 4.2 sCMOS camera (Andor). All fluorescence images were deconvolved in SoftWoRx 5.0 software (Applied Precision) with lens-specific PSFs. Post-fixation, embedding, and sectioning were done as previously described[32]. Serial thin sections (70 nm) were imaged on a JEOL 1400 microscope operated at 80 kV using a side-mounted 4.0 Megapixel XR401 sCMOS camera (Advanced Microscopy Techniques Corp) controlled by AMT Capture Engine ver.7.0.

For CLEM of interphase MN (Extended Data Fig. 5g, four cells with three intact MN and four newly disrupted MN), RPE-1 cells expressing H2B–eGFP and RFP–NLS were synchronized by the nocodazole block (100 ng ml^{-1}) and release protocol and replated as described above. Prior to live-cell imaging, multiple *xyz* positions for subsequent analysis were marked with a diamond scribe. Images were acquired at 10-min intervals and micronuclear disruption was detected by abrupt loss of RFP–NLS. Cells selected for electron microscopy were then fixed and processed as described above. Note: similar results were obtained from the analysis of four cells (five intact and two disrupted MN) that were collected from fixed samples (without live-cell imaging) 16 h after release from the nocodazole block (not shown).

Structured illumination microscopy (SIM). For 3D-SIM of RPE-1 cells expressing emerin–eGFP (Extended Data Fig. 5f) or HeLa K cells expressing CHMPB–eGFP (Extended Data Fig. 8a–c), lagging chromosomes were induced by a nocodozole block and release protocol. After nocodazole release, cells were replated onto coverslips (MatTek, high tolerance coverslips) that were pre-coated with 0.1 mg ml^{-1} poly-D-lysine hydrobromide (Sigma-Aldrich). To obtain anaphase/telophase cells, cells were fixed during mitotic exit (~50 min or 90 min post nocodazole release for RPE-1 and HeLa K cells, respectively) in 4% EM grade formaldehyde (Polysciences) diluted in PEM buffer (80 mM K-Pipes, 5 mM EGTA, 1 mM MgCl2 (pH 6.8)) for 5 min at 37 °C. Cells were then permeabilized with PEM-0.15% Triton X-100 for 2 min and processed for immunofluorescence. Primary and secondary antibodies were diluted in PBS-0.05% saponin and incubated for 2–3 h. To minimize photobleaching of samples, CHMP4B–eGFP or emerin–eGFP signal was enhanced by labelling with FluoTag (Atto488)-X4 anti-GFP (1:250, NanoTag Biotechnologies).

SIM was performed on an OMX V4 Blaze (GE Healthcare) microscope equipped with three watercooled PCO.edge sCMOS cameras, and 405-nm, 488-nm, 568-nm and 642-nm laser lines. Images were acquired with a 0.125-μm step size without binning using a ×60 1.42 NA Plan Apochromat objective (Olympus). For each *z*-section, 15 raw images (three rotations with five phases each) were acquired. To minimize the photobleaching of GFP in thick mitotic specimens, *z*-stack collection was limited by manual selection to the middle (3.5–5 μm) region of the cell that contained the entire lagging chromosome. Images were acquired from ~28–40 *z*-sections with a 0.125-μm step size. Spherical aberration was minimized by immersion oil matching[33]. Super-resolution images were computationally reconstructed from the raw data sets with a channel-specific, measured optical transfer function and a Wiener filter constant of 0.001 using CUDA-accelerated 3D-SIM reconstruction code based on Gustafsson et al. (2008)[34]. TetraSpeck beads (Thermo Fisher) or a nano-grid control slide (GE) were used to measure axial and lateral chromatic misregistration, and experimental data sets were registered using the imwarp function in MATLAB (MathWorks). 3D rendered images were generated using Imaris image analysis software (Bitplane).

Statistical analysis. Statistical analysis was performed using Prism (GraphPad software Inc.). Student's *t*-test or non-parametric tests (Mann–Whitney test) were performed on the basis of sample distributions. For bar graphs or scatter dot plots, bars indicate the mean value of population with 95% CI, unless otherwise specified. To determine the correlation between the accumulation of RFP–NLS and RPA2 in MN (Extended Data Fig. 3d, e), Spearman's correlation analysis was used. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

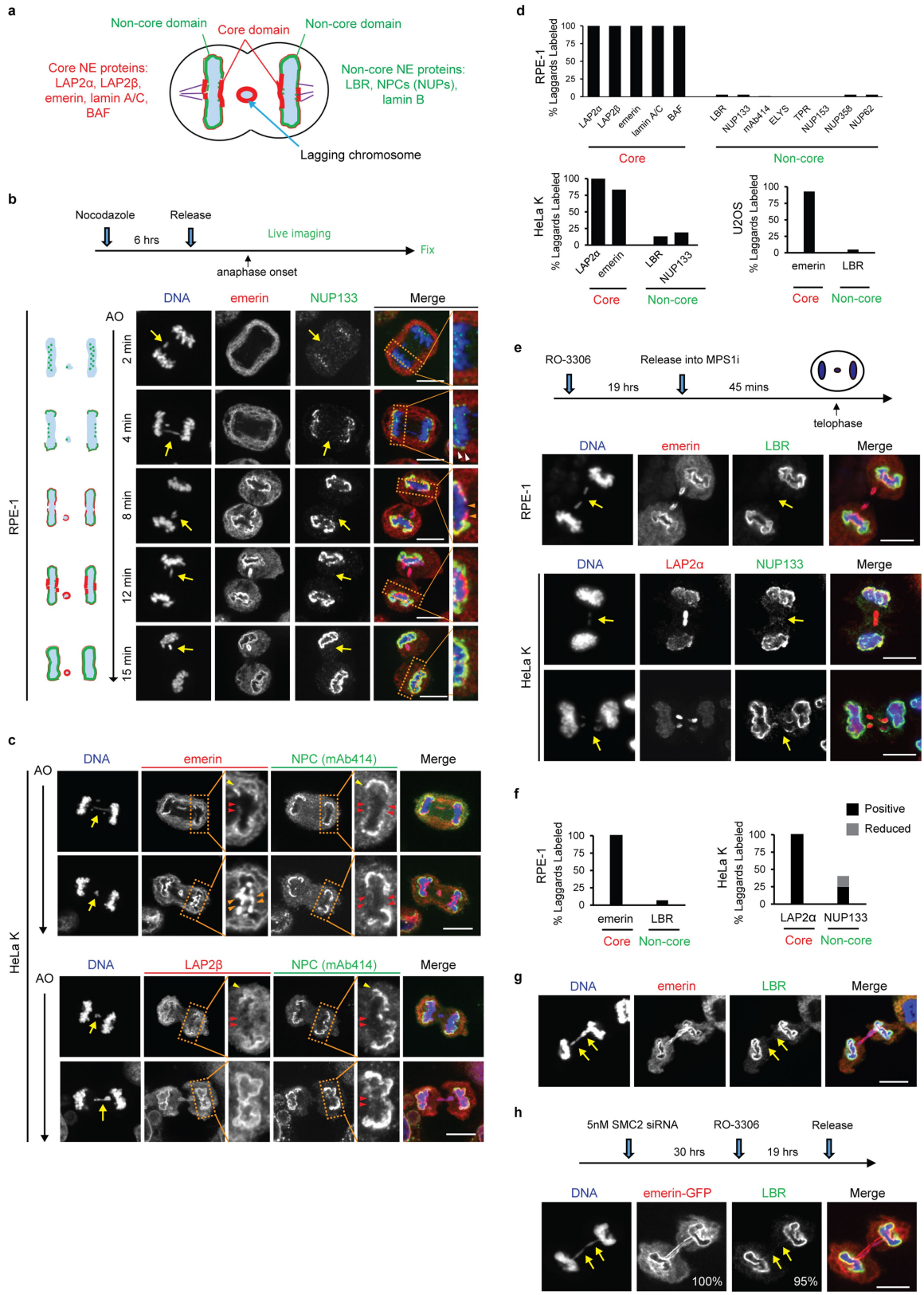
The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files. Source data for all graphs (Figs. 1c, 2a, b, 3a, c, 4b, d, e; Extended Data Figs. 1d, f, 2b, c, d, f, g, 3a–c, e, 4c, d, f, g, 5b, c, e, 6a–c, 7b, d, e, 8d, 9a, d–f, h, 10c, d) are provided with the online version of the paper. All other data sets generated and/or analysed in the current study are available from the corresponding author upon reasonable request. Supplementary Fig. 1 contains scanned complete images of western blots.

29. Poser, I. et al. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409–415 (2008).
30. Gruneberg, U., Neef, R., Honda, R., Nigg, E. A. & Barr, F. A. Relocation of Aurora B from centromeres to the central spindle at the metaphase to anaphase transition requires MKlp2. *J. Cell Biol.* **166**, 167–172 (2004).
31. Soto, M. et al. p53 prohibits propagation of chromosome segregation errors that produce structural aneuploidies. *Cell Reports* **19**, 2423–2431 (2017).
32. Rieder, C. L. & Cassels, G. Correlative light and electron microscopy of mitotic cells in monolayer cultures. *Methods Cell Biol.* **61**, 297–315 (1999).
33. Hiraoka, Y., Sedat, J. W. & Agard, D. A. Determination of three-dimensional imaging properties of a light microscope system. Partial confocal behavior in epifluorescence microscopy. *Biophys. J.* **57**, 325–333 (1990).
34. Gustafsson, M. G. et al. Three-dimensional resolution doubling in wide-field fluorescence microscopy by structured illumination. *Biophys. J.* **94**, 4957–4970 (2008).

© 2018 Springer Nature Limited. All rights reserved.

35. Belgareh, N. et al. An evolutionarily conserved NPC subcomplex, which redistributes in part to kinetochores in mammalian cells. *J. Cell Biol*. **154**, 1147–1160 (2001). https://doi.org/10.1083/jcb.200101081.
36. Maeshima, K. et al. Cell-cycle-dependent dynamics of nuclear pores: pore-free islands and lamins. *J. Cell Sci*. **119**, 4442–4451 (2006).
37. Clever, M., Funakoshi, T., Mimura, Y., Takagi, M. & Imamoto, N. The nucleoporin ELYS/Mel28 regulates nuclear envelope subdomain formation in HeLa cells. *Nucleus* **3**, 187–199 (2012).
38. Mimura, Y., Takagi, M., Clever, M. & Imamoto, N. ELYS regulates the localization of LBR by modulating its phosphorylation state. *J. Cell Sci*. **129**, 4200–4212 (2016).
39. Hudson, D. F., Vagnarelli, P., Gassmann, R. & Earnshaw, W. C. Condensin is required for nonhistone protein assembly and structural integrity of vertebrate mitotic chromosomes. *Dev. Cell* **5**, 323–336 (2003).
40. Maiato, H., Afonso, O. & Matos, I. A chromosome separation checkpoint: A midzone Aurora B gradient mediates a chromosome separation checkpoint that regulates the anaphase-telophase transition. *BioEssays* **37**, 257–266 (2015).
41. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and kataegis induced by telomere crisis. *Cell* **163**, 1641–1654 (2015).
42. Denais, C. M. et al. Nuclear envelope rupture and repair during cancer cell migration. *Science* **352**, 353–358 (2016).
43. Steigemann, P. et al. Aurora B-mediated abscission checkpoint protects against tetraploidization. *Cell* **136**, 473–484 (2009).
44. Uehara, R. et al. Aurora B and Kif2A control microtubule length for assembly of a functional central spindle during anaphase. *J. Cell Biol*. **202**, 623–636 (2013).
45. Sagona, A. P., Nezis, I. P. & Stenmark, H. Association of CHMP4B and autophagy with micronuclei: implications for cataract formation. *BioMed Res. Int.* **2014**, 974393 (2014).
46. Santaguida, S. et al. Chromosome mis-segregation generates cell-cycle-arrested cells with complex karyotypes that are eliminated by the immune system. *Dev. Cell* **41**, 638–651 (2017).
47. Shah, P., Wolf, K. & Lammerding, J. Bursting the bubble—nuclear envelope rupture as a path to genomic instability? *Trends Cell Biol*. **27**, 546–555 (2017).

© 2018 Springer Nature Limited. All rights reserved.

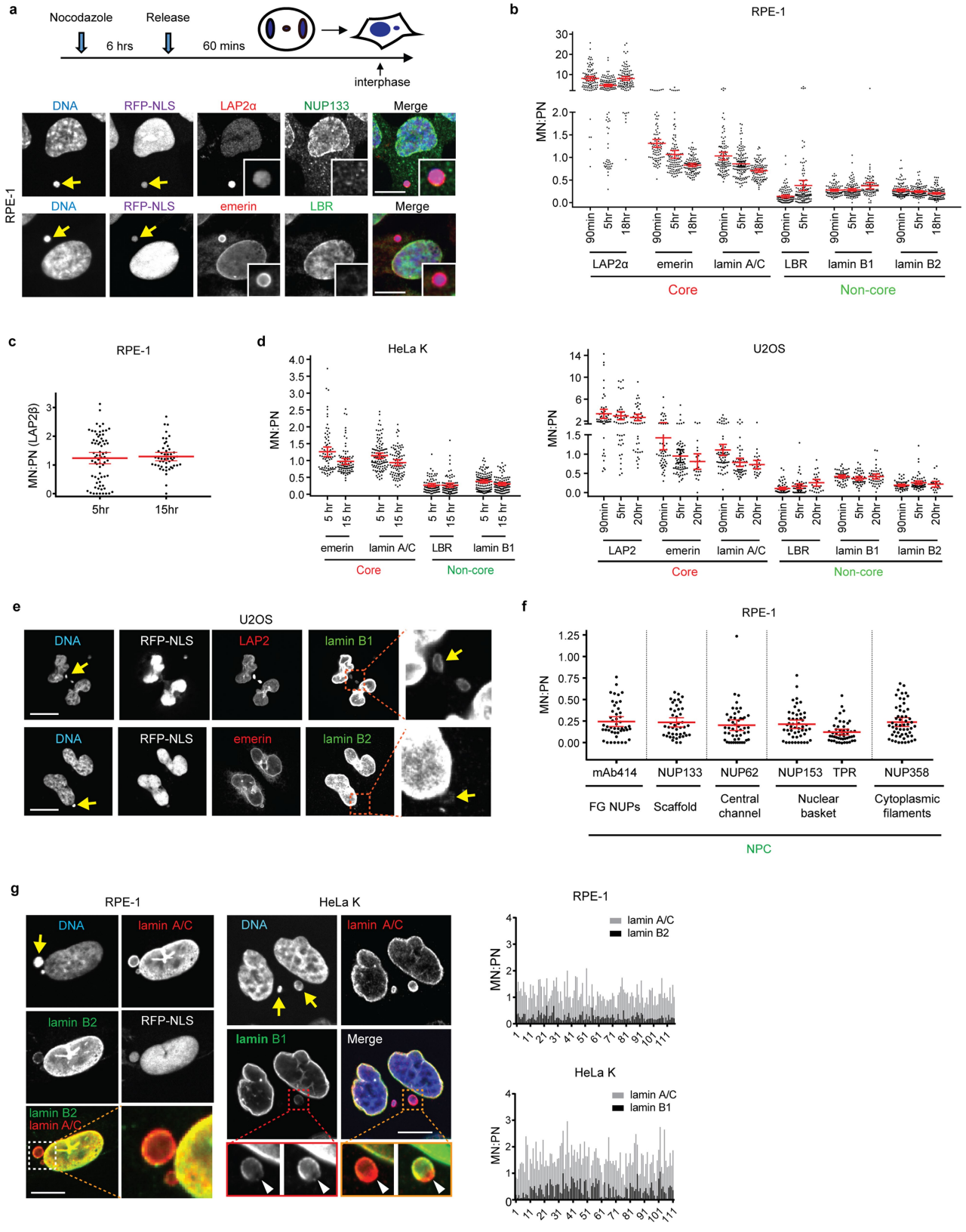


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 1 | Defective NE assembly on lagging chromosomes (or chromosome bridges). a, Cartoon of NE subdomains that form transiently around the main chromosome mass in a telophase cell[8,9] and summarizing results for lagging chromosomes. DNA, light blue. BAF, barrier-to-autointegration factor. The NPC is a complex containing NUPs, including NUP133, NUP107, ELYS, TPR, NUP153, NUP358 and NUP62[1] (see **d** and Extended Data Fig. 2f). **b**, Live-cell/fixed cell imaging showing the recruitment of core (emerin) and non-core (NUP133) proteins onto lagging chromosomes (yellow arrows) and the main chromosome mass at the indicated times after anaphase onset (AO, $t = 0$). Top, experimental scheme. RFP–H2B-expressing RPE-1 cells were released from the nocodazole mitotic block, imaged on gridded coverslips at 2-min intervals, and then fixed and labelled for immunofluorescence (Extended Data Fig. 4e). Bottom, cartoons (left) and images of RPE-1 cells (right) showing DNA (blue), emerin (red) and NUP133 (green). Each image represents 20 cells (from two technical replicates) from the indicated time points. Scale bars, 10 μm. Note there is some variability between cells (~1–2 min) in the times at which different proteins are recruited. During early–mid anaphase (2–4 min post AO), NUP133 is on kinetochores[35] (green dots in cartoon). About 4–6 min after AO, NUP133 begins to assemble on the chromosome periphery (right-most column, white arrowheads). About 6–8 min after AO, emerin assembles on lagging chromosome and the main chromosome mass, including the region adjacent to the central spindle (right-most column, orange arrowheads, microtubules not shown). By 8–12 min after AO, emerin becomes concentrated in a recognizable core domain[8], which is also detected as a gap in NUP133 signal (enlarged image). The peripheral localization of NUP133 (for example, 8 min) marks the non-core domain. About half of the NPCs of the interphase nucleus assemble in this ~8–10-min period of telophase[28] (hereafter referred to as late mitotic NPC assembly). Nuclear pore proteins display the most obvious non-core gap, whereas other non-core proteins, such as LBR, more commonly display reduced signal intensity within the core domain (**e**, top row, RPE-1). Defective NUP133 assembly on lagging chromosomes persists throughout mitotic exit (>15 min after AO). By ~15 min after AO, the core and non-core domains become intermingled on the main nucleus, with fragments of the core domain persisting as pore-free islands that are slowly populated by NPCs during interphase[10,36]. **c**, Similar to RPE-1 cells (**b**), images of HeLa K cells (representative of 30 cells, from two experiments) showing that the core membrane proteins emerin (top two rows) and LAP2β (bottom two rows) first associate with the chromosome periphery (yellow arrowheads) contemporaneously with the non-core (NPC, mAb414 detects FG-containing nucleoporins) proteins. About 2–4 min later, the core proteins extend into and then concentrate in (emerin, orange arrowheads; LAP2β does not concentrate) the core domain (red arrowheads). Cells were synchronized as in **e**. In HeLa K cells, lagging chromosomes often exhibit a slight delay (~1–2 min) in the recruitment of core membrane proteins (emerin and LAP2β) as compared to the periphery of the main chromosome mass. Scale bars, 10 μm. It is thought that the NE assembles from a continuous network of mitotic endoplasmic reticulum (ER)[1]. It is therefore simplest to propose that the core and non-core subdomains are also in a continuous network, but the core domain is just a region of the continuous ER network that is missing the non-core subgroup of proteins. Supporting this idea, prior work[8] and our data (**b**, **c**) show that the core proteins initially assemble together with the non-core proteins around the chromosome periphery and only later become enriched near the microtubules. This data suggests that the domain partitioning could come solely from NPC precursors and LBR (which requires ELYS for recruitment[37,38]) being preferentially retained in fenestrated ER sheets[28] that might less readily penetrate bundled spindle microtubules (Fig. 3a). Although we favour this model, we cannot exclude the possibility that core and non-core proteins somehow partition into separate membrane compartments. **d**, Quantification of defective non-core NE protein recruitment to lagging chromosomes. Synchronization as in Fig. 1a (n = 64, 118, 151, 124, 151, 145, 150, 69, 90, 65, 70, 60, 64, left to right, from three experiments for RPE-1 cells, $n = 149$, 110, 124, 132, from two experiments for HeLa K cells and $n = 44$, 76, from two experiments for U2OS cells). **e**, **f**, Orthogonal method (1 μM NMS-P715, MPS1i) to generate lagging chromosomes shows a similar non-core NE assembly defect to the nocodazole block-and-release protocol (Fig. 1a). **e**, Top, experimental scheme. Bottom, representative images of RPE-1 and HeLa K cells. In RPE-1 and U2OS cells there is a near absence of non-core protein on lagging chromosomes, regardless of the method of generation. In HeLa K cells, the effect is less penetrant. About 60% of lagging chromosomes lack detectable non-core protein recruitment, ~15% display strongly reduced levels (scored as labelled but shown in grey bar as reduced, **f**), and ~25% display a clear signal (NUP133), but often only covering part of the circumference of the lagging chromosome. These differences are probably due to differences in spindle organization between cell lines (Fig. 3; Extended Data Figs. 6–8). **f**, Quantification of the results ($n = 56$, 78 for RPE-1 cells, from two experiments; $n = 75$, 174 for HeLa K cells, from three experiments). Scale bars, 10 μm. **g**, **h**, Chromatin bridges (arrows) formed after nocodazole release or partial depletion of SMC2[39] show core-only NE protein assembly. **g**, An RPE-1 cell after release from a nocodazole block (representative of 30 DNA bridges from five experiments). **h**, Top, experimental scheme for generating chromosome bridges by partial SMC2 depletion. Bottom, an emerin–GFP-expressing RPE-1 cell. Percentages of cells with the indicated staining pattern are shown ($n = 30$, from two experiments). Scale bars, 10 μm. DNA bridges are uniformly depleted for non-core (LBR) proteins, with no evidence for a gradient as might be expected for the chromosome separation checkpoint hypothesis[15,40]. Interphase chromatin bridges have been reported to have an altered NE protein composition, including reduced levels of lamin B1 and NPCs[41], which is consistent with our findings.

© 2018 Springer Nature Limited. All rights reserved.

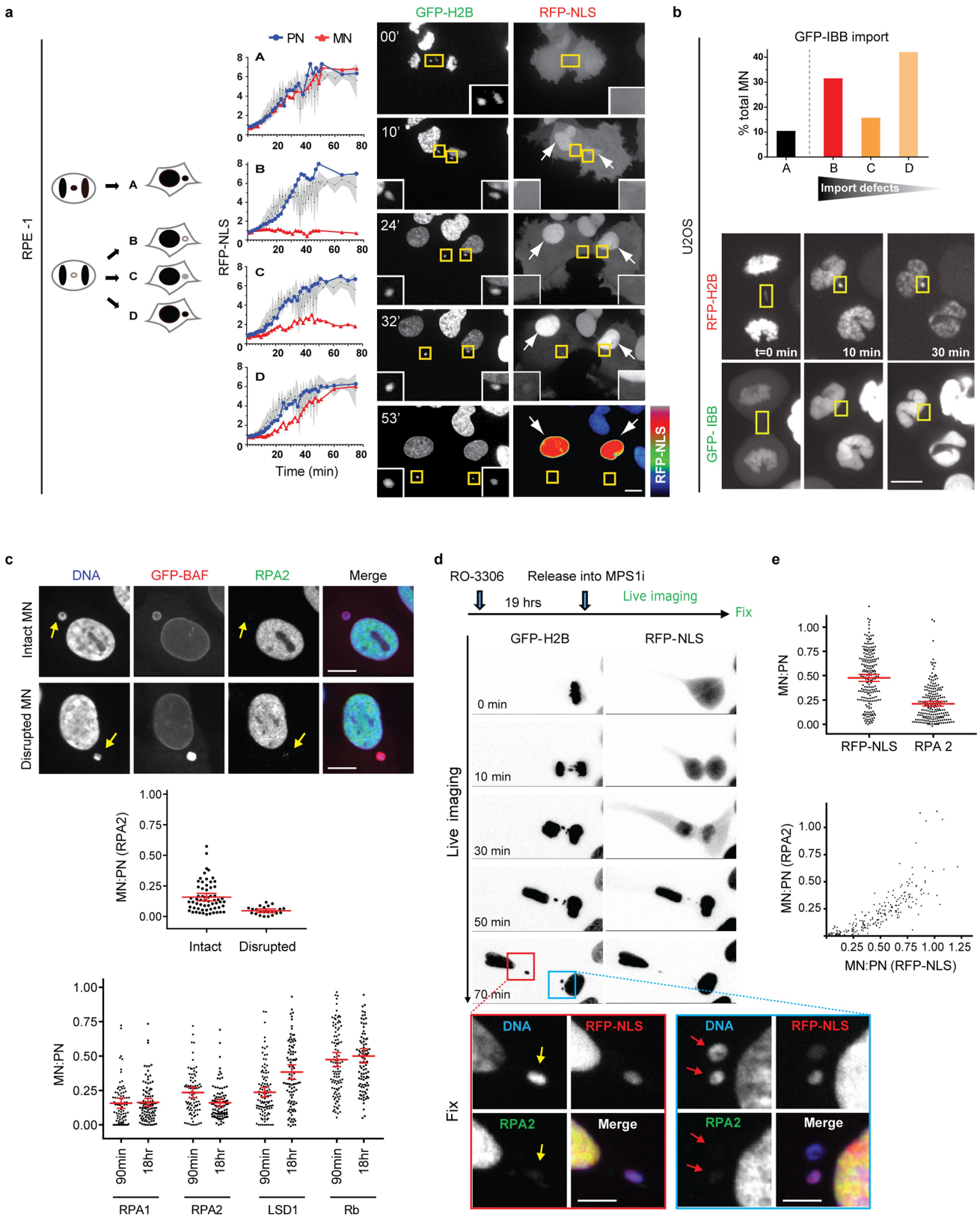


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 2 | Defect in NPC and other non-core protein assembly persists in interphase MN in multiple cell lines. **a**–**c**, Defective non-core NE protein recruitment to MN in RPE-1 cells. **a**, Top, experimental scheme. Bottom, representative images (quantified in **b**) with MN (arrows in enlarged images). Red letters, core NE proteins; green, non-core proteins. **b**, Quantification of results from **a**. The FI ratio of the indicated NE proteins in intact MN (RFP–NLS positive) relative to PN at the indicated time points after release from nocodazole block (mean with 95% CI, $n = 97, 120, 106, 92, 114, 104, 113, 116, 104, 114, 121, 91, 114, 89, 113, 116, 125$, left to right, from two experiments). Scale bars, 10 μm. **c**, MN/PN FI ratio of LAP2β in RPE-1 cells at the indicated time points after release from nocodazole block (mean with 95% CI, $n = 70, 49$, from two experiments). **d**, Deficiency of non-core proteins persists in MN in HeLa K (left) and U2OS cells (right). MN/PN FI ratio of the indicated NE proteins after release from nocodazole block as in **a** (mean with 95% CI, $n = 79, 84, 111, 90, 79, 84, 111, 90$, left to right for HeLa K, from two experiments; $n = 53, 47, 45, 53, 66, 23, 64, 53, 27, 66, 54, 28, 53, 47, 46, 53, 69, 23$, left to right for U2OS, from two experiments). **e**, Representative images of U2OS cells from **d** showing reduced assembly of B-type lamins in MN (arrows) 90 min after nocodazole release. Scale bars, 10 μm. **f**, Reduced NPC assembly on MN. MN/PN FI ratio of NPC proteins in RPE-1 cells 15 h after release from nocodazole block (mean with 95% CI, $n = 47, 44, 49, 51, 49, 56$, left to right, from two experiments). **g**, Reduced accumulation of B-type lamins but not A-type lamins on intact MN. Left, representative images of RPE-1 (left) and HeLa K (right) cells. In the enlarged merged images from the orange boxed region (HeLa K), the intensity of lamin B1 has been scaled differently to illustrate two points. First, there is reduction of lamin B but not lamin A/C in MN, which becomes evident when lamin B and lamin A/C on the main nucleus are scaled to the same intensity. Second, some MN display a lamin B1 gap, as has been reported[6] (arrowheads indicate lamin B1 gap formed where the lamin A/C rim is continuous), which becomes evident when the lamin B1 intensity is scaled to a higher level. Right, general reduction of B-type lamins in MN. MN/PN FI ratio for the indicated lamins. Shown is the background-subtracted raw data for each individual micronucleus analysed in the indicated cell lines ($n = 116$ for RPE-1, $n = 111$ for HeLa K, from **b**, **d** above), 5 h after nocodazole release. Scale bars, 10 μm. A fraction of MN exhibit reduced lamin A/C, possibly owing to impaired import of lamin A/C during interphase.

© 2018 Springer Nature Limited. All rights reserved.

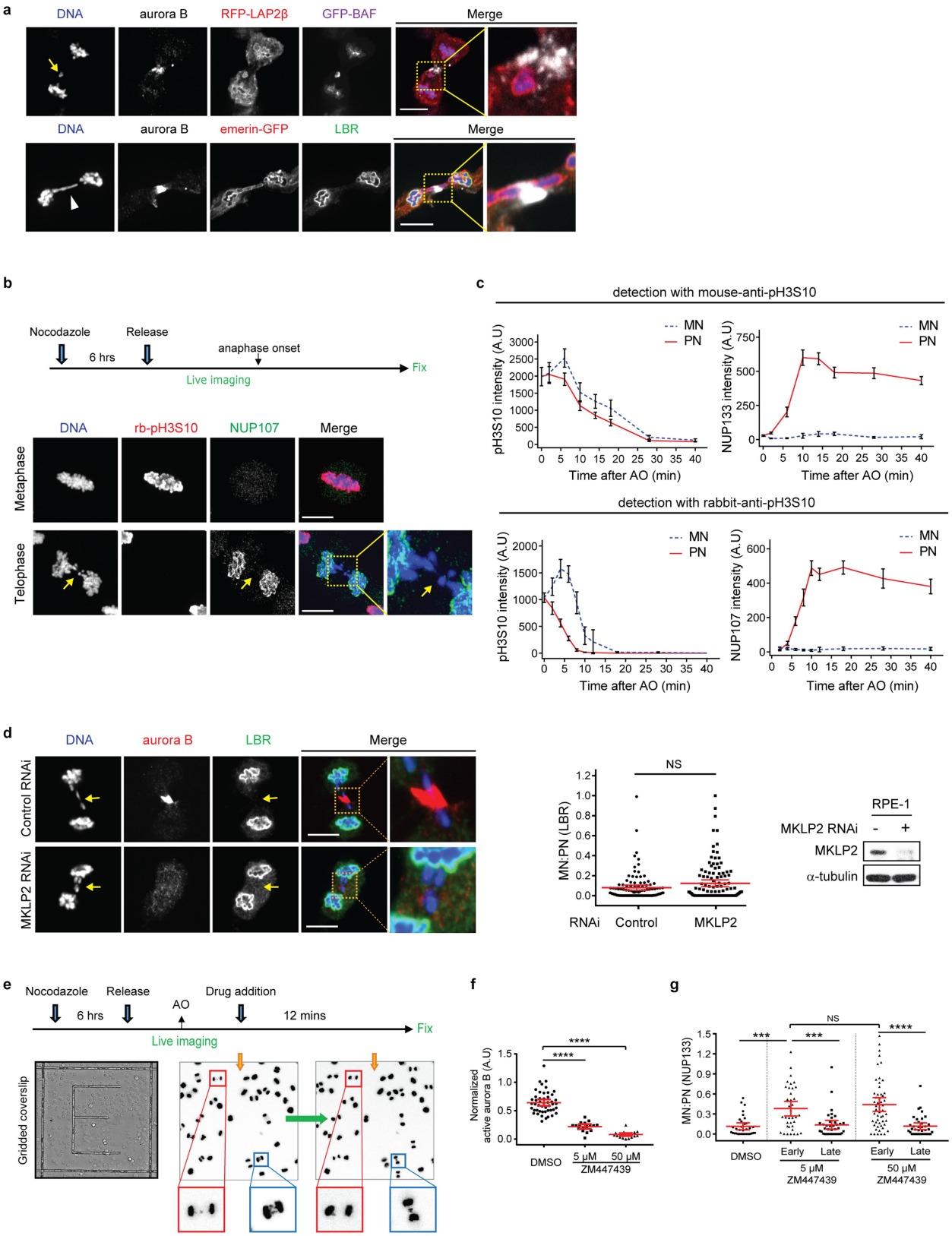


Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 3 | MN have defective nuclear import and impaired accumulation of nuclear proteins. a, **b**, Defective nuclear import kinetics in newly formed MN visualized by high temporal resolution, confocal live-cell imaging. MN were generated in RPE-1 (**a**) or U2OS (**b**) cells as in Fig. 1a. **a**, Defective nuclear RFP–NLS import in RPE-1 cells. Left, categories of cells with differing levels of accumulation of the import reporter in MN as compared to PN (see Fig. 1c for quantification). Middle, nuclear import kinetics from representative RPE-1 cells expressing GFP–H2B and RFP–NLS for PN (blue) and MN (red). Black lines and shaded regions indicate mean ± s.d. from 11 PN. To control for photobleaching, the bleaching coefficient of the RFP–NLS signal was experimentally determined by imaging of cells during late interphase when nuclear import was presumed to be at steady state. A correction based on this bleaching coefficient obtained for 12 PN/MN pairs was then applied to the data presented here (see Methods). Right, representative images from a time-lapse series from RPE-1 cells expressing GFP–H2B (left) and RFP–NLS (right). Arrows indicate a pair of PN with RFP–NLS import evident from 10 min. Yellow boxes show lagging chromosomes and the resulting MN with nuclear RFP–NLS accumulation defects (insets, enlarged images). Bottom right, heat map of RFP–NLS intensity ($t = 53$ min) illustrating the strong import defect in MN (see Supplementary Videos 1, 2). **b**, Defective nuclear GFP–IBB import in lagging chromosomes from U2OS cells. Top, percentage of MN corresponding to the categories in **a** in U2OS cells expressing RFP–H2B and GFP–IBB ($n = 19$, from 17 experiments). Bottom, images from a time-lapse series (boxes, lagging chromosomes and MN). Scale bars, 10 μm. **c**, Impaired nuclear accumulation of multiple proteins in MN relative to PN. Top, representative images of RPE-1 cells with intact or disrupted MN (arrows). Micronucleus disruption is visualized by hyper-accumulation of GFP–BAF[42]. Both intact and disrupted MN lack detectable accumulation of RPA2. Middle, MN/PN FI ratio of RPA2 (mean with 95% CI, $n = 62, 23$, from two technical replicates). MN were generated in RPE-1 cells as in Extended Data Fig. 1e. Bottom, FI ratio of indicated proteins in intact (RFP–NLS positive) MN from RPE-1 cells at indicated time points after nocodazole release as in Extended Data Fig. 2a (mean with 95% CI, $n = 89, 98, 89, 98, 107, 100, 107, 100$, left to right, from three experiments). LSD1, lysine-specific histone demethylase 1; Rb, retinoblastoma protein. Scale bars, 10 μm. **d**, **e**, Live-cell/fixed cell imaging showing impaired nuclear accumulation of RPA2 and RFP–NLS in MN after mitotic exit. **d**, Top, experimental scheme. Bottom, images from wide-field time-lapse series followed by fixation and confocal imaging. Confocal images (from red and blue boxed regions) illustrate reduced accumulation of RFP–NLS and RPA2 in newly formed MN (yellow and red arrows, respectively). Time after anaphase onset ($t = 0$ min) is shown. Scale bars, 10 μm. **e**, Top, MN/PN FI ratio of RPA2 and RFP–NLS in RPE-1 cells ~1 h after anaphase onset (mean with 95% CI, $n = 211$, from two experiments). Bottom, defective accumulation of RPA2 and RFP–NLS are correlated ($r = 0.9039$, $P < 0.0001$, two-tailed Spearman's correlation).

© 2018 Springer Nature Limited. All rights reserved.

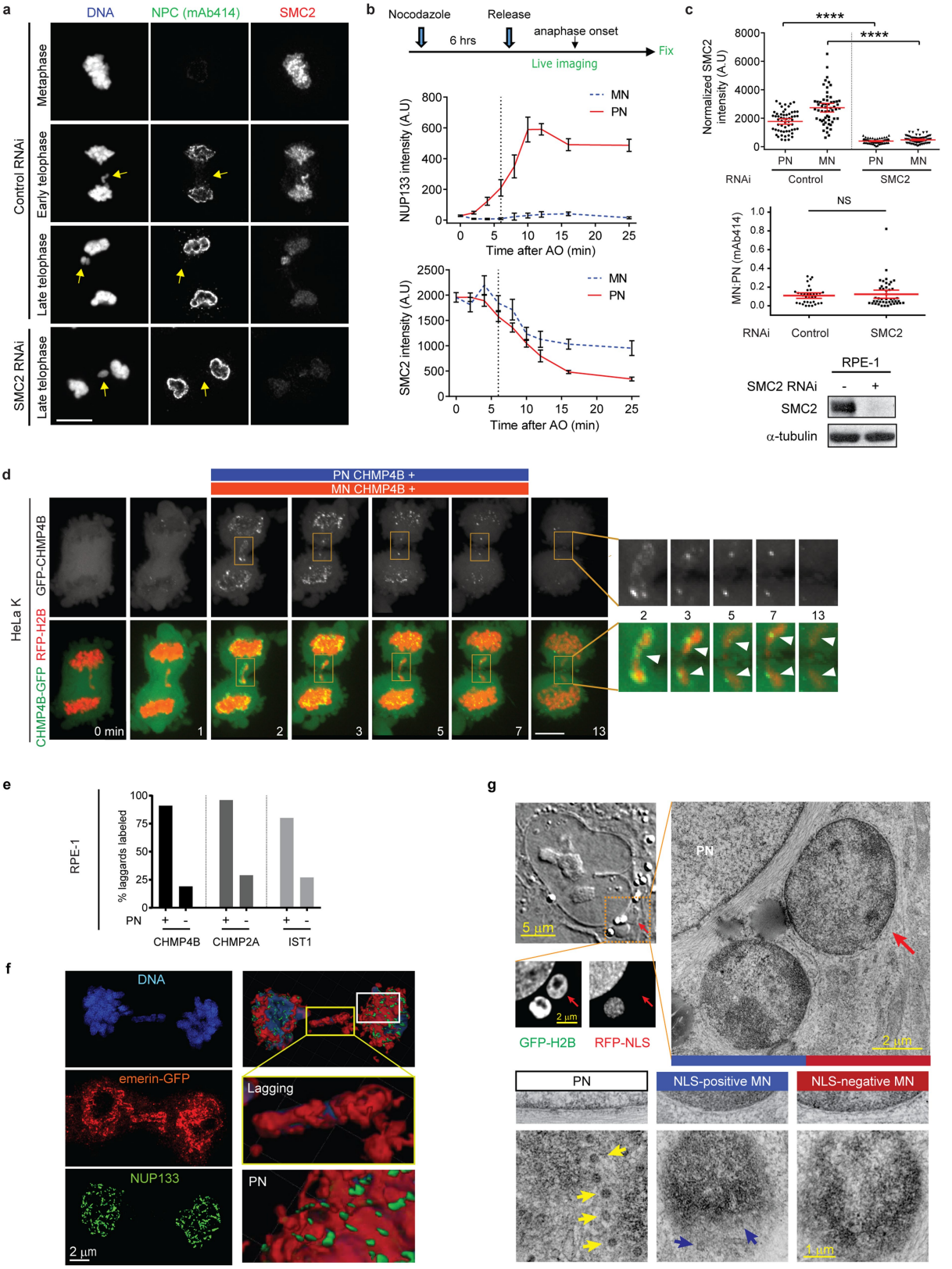


Extended Data Fig. 4 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 4 | Proximity to spindle midzone Aurora B does not affect the NE assembly defect of lagging chromosomes; this defect persists through late telophase regardless of whether Aurora B is active. **a**, Core membrane proteins (LAP2β and emerin) assemble on lagging chromosomes and chromosome bridges, independent of their position relative to the spindle midzone. Top, images of an RPE-1 cell expressing GFP–BAF and RFP–LAP2β show recruitment of LAP2β to lagging chromosome (arrow). Bottom, images of an RPE-1 cell expressing emerin–GFP show recruitment of emerin to a chromosome bridge (arrowhead). Both LAP2β and emerin can assemble near the concentration of aurora B (white, in merged images) at the spindle midzone (representative of 30 cells, from two experiments). Cells synchronized as in Fig. 1a. Scale bars, 10 μm. **b**, **c**, Live-cell/fixed cell imaging shows that loss of H3S10 phosphorylation in late telophase does not enable non-core NE assembly onto lagging chromosomes. **b**, Top, experimental scheme. Bottom, representative images of RPE-1 cells. Scale bars, 10 μm. **c**, FI measurements from the main chromosome mass (PN) and lagging chromosome (MN) showing loss of H3S10 phosphorylation relative to the recruitment of NPC proteins in the same cell at indicated time points. Top, phospho-H3S10 with a mouse monoclonal anti-pH3S10 antibody (mean with 95% CI, $n = 31, 46, 34, 27, 26, 28, 34, 26$ each time point for MN and PN mouse-pH3S10, from two experiments; $n = 31, 46, 34, 27, 26, 28, 34, 26$ each time point for MN and PN NUP133, from two experiments). Bottom: rabbit polyclonal anti-pH3S10 antibody (mean with 95% CI, $n = 27, 20, 33, 29, 22, 30, 22, 32, 31, 27$ each time point for MN and PN rabbit-pH3S10, from two experiments; $n = 30, 49, 28, 32, 30, 33, 32, 32, 27$ each time point for MN and PN NUP107, from two experiments). Zero time point measurements for both MN and PN were from metaphase chromosomes. The specificity of both anti-pH3S10 antibodies was confirmed by the near-complete loss of labelling after treatment of nocodazole-arrested mitotic cells with 50 μM ZM447439 (not shown). Note that NUP133 starts to assemble on the main chromosomes (~4–6 min after anaphase onset[28], also see Extended Data Fig. 5b) with readily detected H3S10 phosphorylation. **d**, MKLP2 knockdown, disrupting the transport of aurora B to the spindle midzone, fails to restore non-core (LBR) NE assembly to lagging chromosomes (arrows). Synchronization as in Fig. 1a. Left, representative images of control or MKLP2-depleted RPE-1 cells. Middle, MN/PN FI ratio of LBR. For MN in MKLP2-depleted cells, only cells lacking detectable midzone aurora B were quantified (mean with 95% CI, $n = 118, 105$, from three experiments). NS, $P = 0.8147$, two-tailed Mann–Whitney test. Right, western blot showing MKLP2 knockdown in RPE-1 cells (two experiments, for gel source data see Supplementary Fig. 1). Scale bars, 10 μm. **e**, Live-cell/fixed cell imaging protocol. RPE-1 cells expressing RFP–H2B were plated on gridded dishes to identify cells of interest. Cells were imaged at 2-min intervals. Representative images of two live cells (red and blue boxes) upon (left) or after (right, before fixation) ZM447439 addition. **f**, Consistent with prior studies[17,43,44], aurora B inhibition (ZM447439, 5 μM or 50 μM) rapidly (3 min) inactivates the kinase (mean with 95% CI, $n = 48, 17, 19$, from two technical replicates). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. Active aurora B was assessed by the FI ratio of phospho-aurora B (pT232) to total aurora B in the midbody of telophase cells 3 min after addition of ZM447439 or DMSO. RPE-1 cells were synchronized as in Fig. 1a. **g**, Tenfold higher (50 μM) concentration of aurora B inhibitor gives similar results to those obtained with 5 μM inhibitor (Fig. 2a), confirming that early aurora B inhibition (2 min before to 2 min after anaphase onset) but not late (6–12 min after anaphase onset) partially restores non-core protein (NUP133) recruitment to lagging chromosomes. Note that non-core NE is only partially restored on a fraction of lagging chromosomes even with the high dose of inhibitor (early 50 μM ZM447439) even though this treatment led to a uniform loss of phospho-H3S10 on all chromosomes after anaphase entry (not shown). Therefore, the persistent NE assembly defect on lagging chromosomes after aurora B inhibition cannot be explained by residual phospho-H3S10. Instead, the partial effect is explained by the effects of aurora B inhibition on the local organization of microtubules (Extended Data Figs. 6a, 8d). Shown is the MN/PN FI ratio of NUP133 at the indicated time intervals and drug doses (mean with 95% CI, $n = 33, 37, 37, 52, 37$, left to right, from two experiments). Experiments performed as in Fig. 2a in RPE-1 cells. For simplicity, time points corresponding to ± 2 min after anaphase onset were grouped as early treatment, whereas time points corresponding to 6–12 min after anaphase onset were grouped as late treatment. $^{***}P = 0.0001$ (DMSO and early 5 μM ZM), $^{***}P = 0.0003$ (early 5 μM ZM and late 5 μM ZM), $^{****}P < 0.0001$; NS, $P = 0.494$, two-tailed Mann–Whitney test.

© 2018 Springer Nature Limited. All rights reserved.

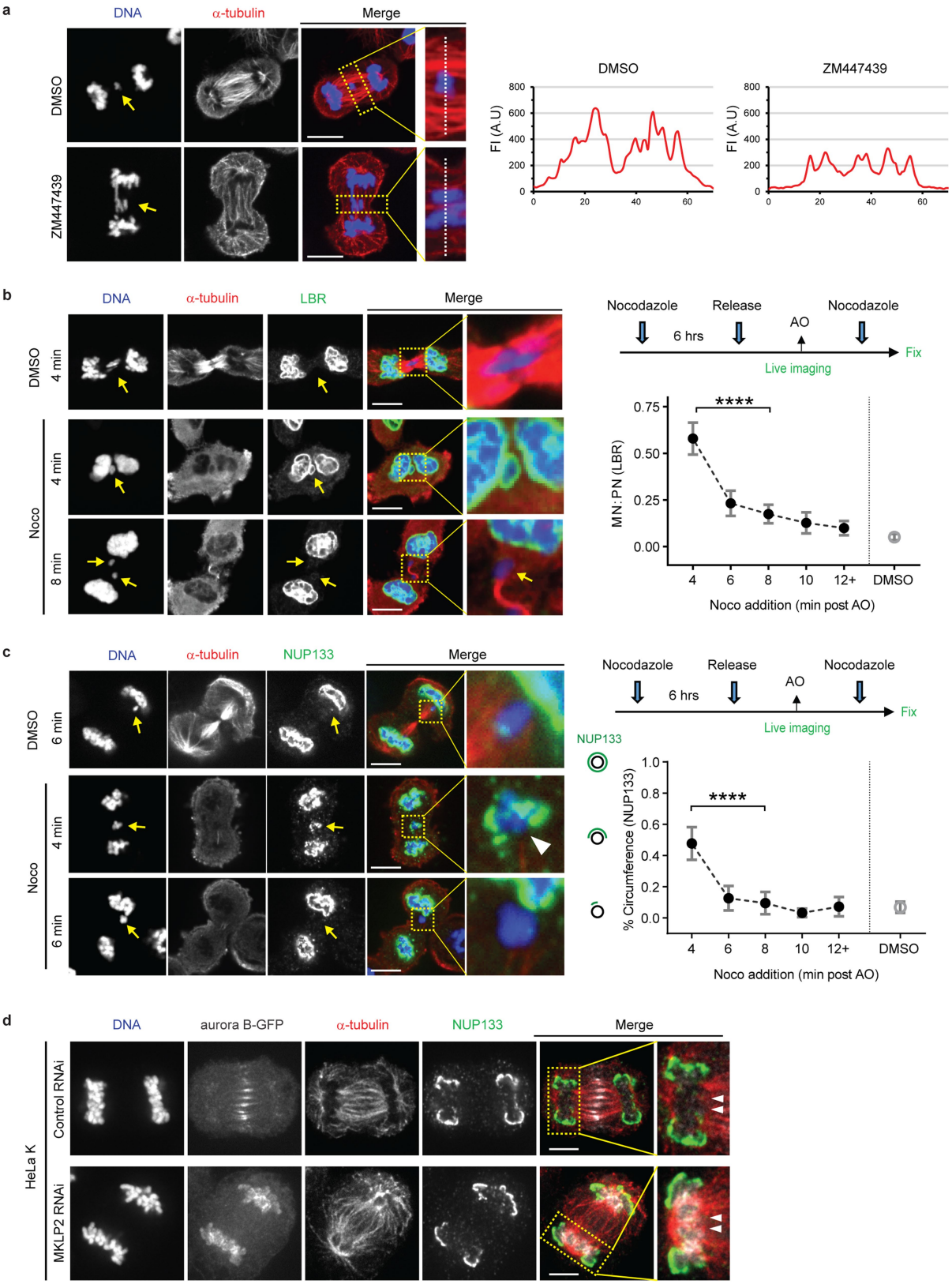


Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 5 | The irreversible NE assembly defect of lagging chromosomes may be caused by enclosure in an NPC-deficient NE, rather than persistent condensins. a–**c**, Loss of condensin (SMC2) is not required for non-core and NPC assembly. **a**, Representative images of control and SMC2-depleted RPE-1 cells showing that loss of SMC2 does not restore non-core assembly on lagging chromosomes (quantified in **b**, **c**). Scale bar, 10 μm. **b**, Top, experimental scheme of live-cell/fixed cell imaging. Middle and bottom, FI of NUP133 and SMC2 on the main chromosome mass (PN) and lagging chromosomes (MN) at time points after anaphase onset in RPE-1 cells (mean with 95% CI, $n = 52$, 37, 35, 44, 35, 36, 32, 46, 41 each timpoint for MN SMC2, from three experiments; $n = 52$, 37, 47, 55, 40, 48, 38, 46, 41 each time point for PN SMC2, from three experiments; $n = 31$, 28, 37, 22, 23, 25, 26, 28, 34 each time point for MN and PN NUP133, from two experiments). Zero time point measurements are from metaphase chromosomes. Dashed lines indicate time points, demonstrating that NUP133 can assemble on the PN when condensins are present at ~70% of their maximum levels in metaphase. Note that for MN (~10 min after anaphase onset), the decline in SMC2 levels is slower than on the PN, presumably because of nuclear transport is required for complete chromosome decondensation. **c**, Top, SMC2 levels on PN and MN during telophase in control and SMC2-depleted cells (mean with 95% CI, $n = 58$, 58, 75, 75, left to right, from two experiments). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. Middle, MN/PN FI ratio of mAb414 to detect nucleoporins. For MN in SMC2-depleted cells, only cells with no detectable SMC2 were quantified for mAb414 (mean with 95% CI, $n = 35$, 45, from two experiments). NS, $P = 0.9251$, two-tailed Mann–Whitney test. Bottom, western blot showing SMC2 knockdown in RPE-1 cells (two experiments; for gel source data see Supplementary Fig. 1). **d**, **e**, Recruitment of ESCRT-III to lagging chromosomes. **d**, Normal kinetics for the association and dissociation of ESCRT-III (CHMP4B–GFP) on lagging chromosomes (white arrowheads) during NE reformation. Lagging chromosomes were induced in HeLa K cells expressing CHMP4B–GFP (green) and RFP–H2B (red) by nocodazole block-and-release as in Fig. 1a and imaged at 1-min intervals on a confocal microscope (Supplementary Video 3). Left, a representative cell showing the CHMP4B residence time on lagging chromosome (MN) as compared to the residence time on the PN. Right, enlarged images of boxed regions showing lagging chromosome (white arrowheads). The duration of CHMP4B association with the PN or MN is shown by the coloured bars above the time-lapse series. Similar results were obtained for chromosome bridges (not shown). Note that prior work suggests that the timely dissociation of CHMP4B is a marker for successful NE sealing and that delayed CHMP4B dissociation is a marker for defective NE sealing[23]. The normal ESCRT-III kinetics on lagging chromosomes or chromosome bridges therefore suggests substantial NE membrane sealing on these structures (representative of ten cells from six experiments). ESCRT-III was not detected on newly formed MN after mitotic exit (not shown). ESCRT-III has been reported to associate with interphase MN[45], which we also observe, but only on disrupted MN (not shown). In addition, the previously described[23] enrichment of ESCRT-III in the core domain of the main chromosome mass is less apparent after nocodazole release (shown here) than is observed when lagging chromosomes are generated by MPS1i treatment (Extended Data Fig. 9b). Scale bars, 10 μm. **e**, ESCRT-III (CHMP4B, CHMP2A and IST1) is recruited to lagging chromosomes in telophase RPE-1 cells. Cells were labelled to detect the indicated proteins synchronized from nocodazole release as in Fig. 1a. Graph shows percentage of lagging chromosomes that were positive for the indicated proteins in early telophase cell (when the main chromosome mass is positive for ESCRT III, PN^+) and late telophase cell (when ESCRT III is no longer detected on the main chromosome mass, PN^-) ($n = 44$, 78, 45, 31, 67, 37, each category, from two experiments). **f**, Near-continuous assembly of core membrane protein around the lagging chromosome. Images of 3D-SIM (left) and Imaris surface renderings (right) of an RPE-1 cell expressing emerin–GFP. Enlarged images show lagging chromosomes (yellow box) and PN (white box): DNA (blue), emerin (red), NUP133 (green) (representative of 22 cells, from two technical replicates). **g**, CLEM showing double-membrane NE (enlarged images in middle panels, cross sections) but reduced NPCs on intact ($RFP–NLS^+$) and newly disrupted MN ($RPF–NLS^-$, red arrows). An RPE-1 cell was fixed <20 min after spontaneous disruption of an MN. Left, top, DIC and fluorescence images immediately after loss of RFP–NLS signal in one of the two MN in the cell. Right, top, 70-nm thin section selected from the full series of the cell. Double-membrane NE is present on the PN as well as on the disrupted (red arrow) and intact MN. Middle row depicts NE at higher magnification. NPCs are prominent in the tangential view (bottom row) of PN (yellow arrows) but not on NLS-positive MN (blue arrows). No NPC is present on the disrupted RPF–NLS-negative MN. Representative of four cells with three intact and four newly disrupted MN, from three experiments (see Methods).

© 2018 Springer Nature Limited. All rights reserved.

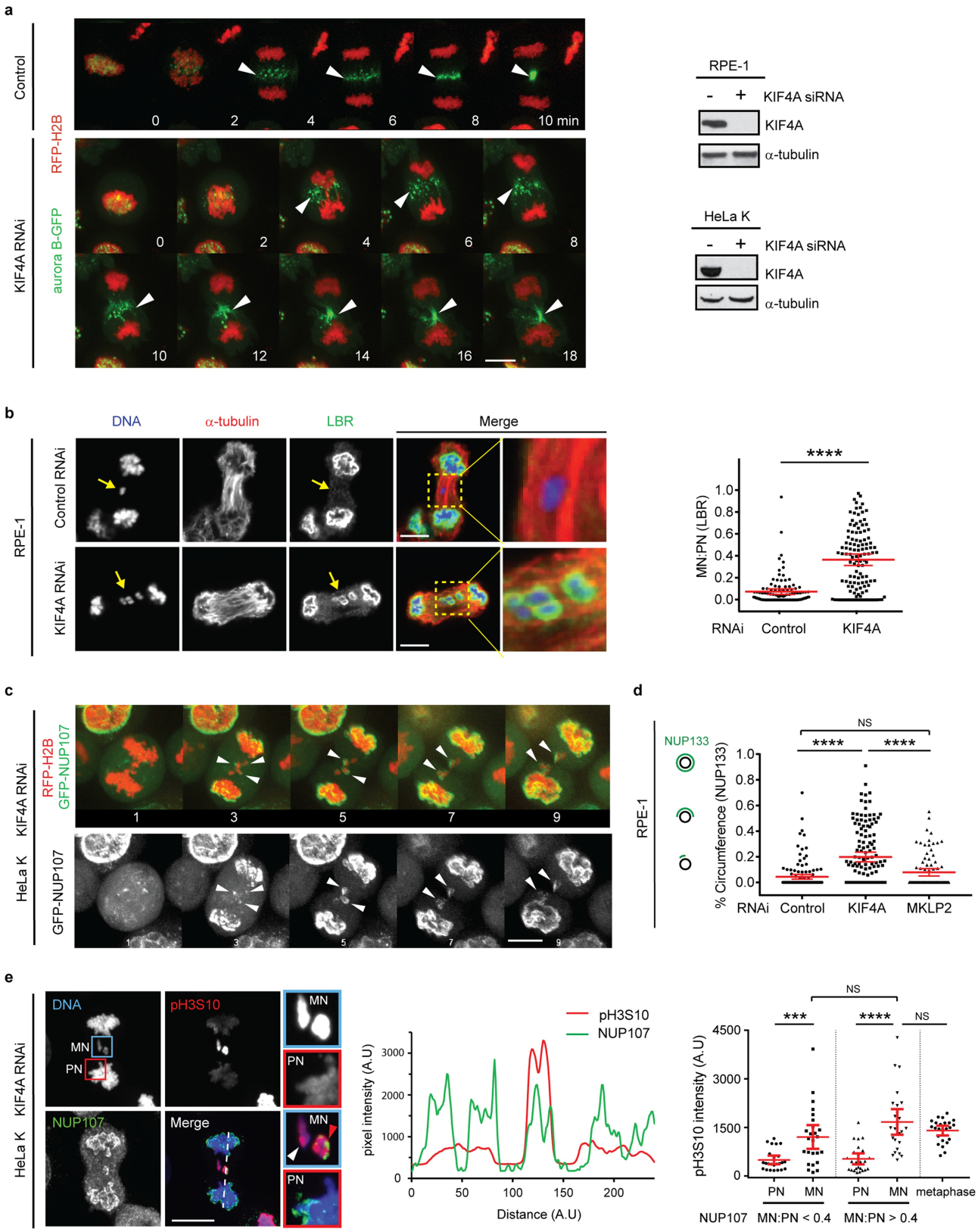


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | Microtubule disruption restores non-core NE assembly to lagging chromosomes only if it occurs early in anaphase. **a**, Inhibition of aurora B (5 μM ZM447439) reduces spindle microtubule bundling and microtubule mass. Left, representative images of RPE-1 cells ($n = 20$, from two technical replicates). Right, linescan profiles of FI for α-tubulin (red) along the white dashed lines (see enlarged images of boxed regions). Scale bars, 10 μm. **b**, **c**, Parallel effects of nocodazole and aurora B inhibitor treatment (Fig. 2a) on non-core NE assembly. Live-cell/fixed cell imaging as in Fig. 2a showing that nocodazole-mediated microtubule depolymerization allows non-core NE assembly only if it occurs early in anaphase. Labelling to detect α-tubulin and LBR is in **b**; labelling to detect α-tubulin and NUP133 is in **c**. Left (**b**, **c**), representative images of RPE-1 cells exposed to DMSO or nocodazole at the indicated times after anaphase onset (cells were fixed and labelled 12 min after DMSO or nocodazole addition). Yellow arrows (**b**, **c**) indicate lagging chromosomes; white arrowhead (**c**) shows that even on a lagging chromosome where NUP133 (green) has largely been restored after nocodazole treatment, the region of this chromosome that remains in contact with residual microtubules (red) is depleted for NUP133. Right (**b**, **c**), experimental scheme (top; Extended Data Fig. 4e) and quantification (bottom). 12+ timepoint on *x*-axes includes cells that were exposed to nocodazole 12–16 min after anaphase onset. **b**, MN/PN FI ratio for LBR (mean with 95% CI, $n = 64, 62, 54, 44, 63, 69$ each time point, from three experiments). **c**, Fraction of lagging chromosome circumference with NUP133 (mean with 95% CI, $n = 33, 29, 26, 25, 28, 68$ each time point, from two experiments). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. Scale bars, 10 μm. **d**, Formation of non-core gap is independent of the normal spatial distribution of aurora B. A non-core protein (NUP133) is excluded from the central spindle region (white arrowheads) on the chromosome mass in HeLa K cells expressing aurora B–GFP (white, in merged images; representative of 30 cells, from two technical replicates). Synchronization as in Extended Data Fig. 1e. NUP133 assembly occurred normally on the periphery of the main chromosome mass, away from the spindle, whether aurora B localized to the central spindle or was forced to remain near the chromosome mass (MKLP2 RNAi)[30]. Scale bars, 10 μm.

© 2018 Springer Nature Limited. All rights reserved.

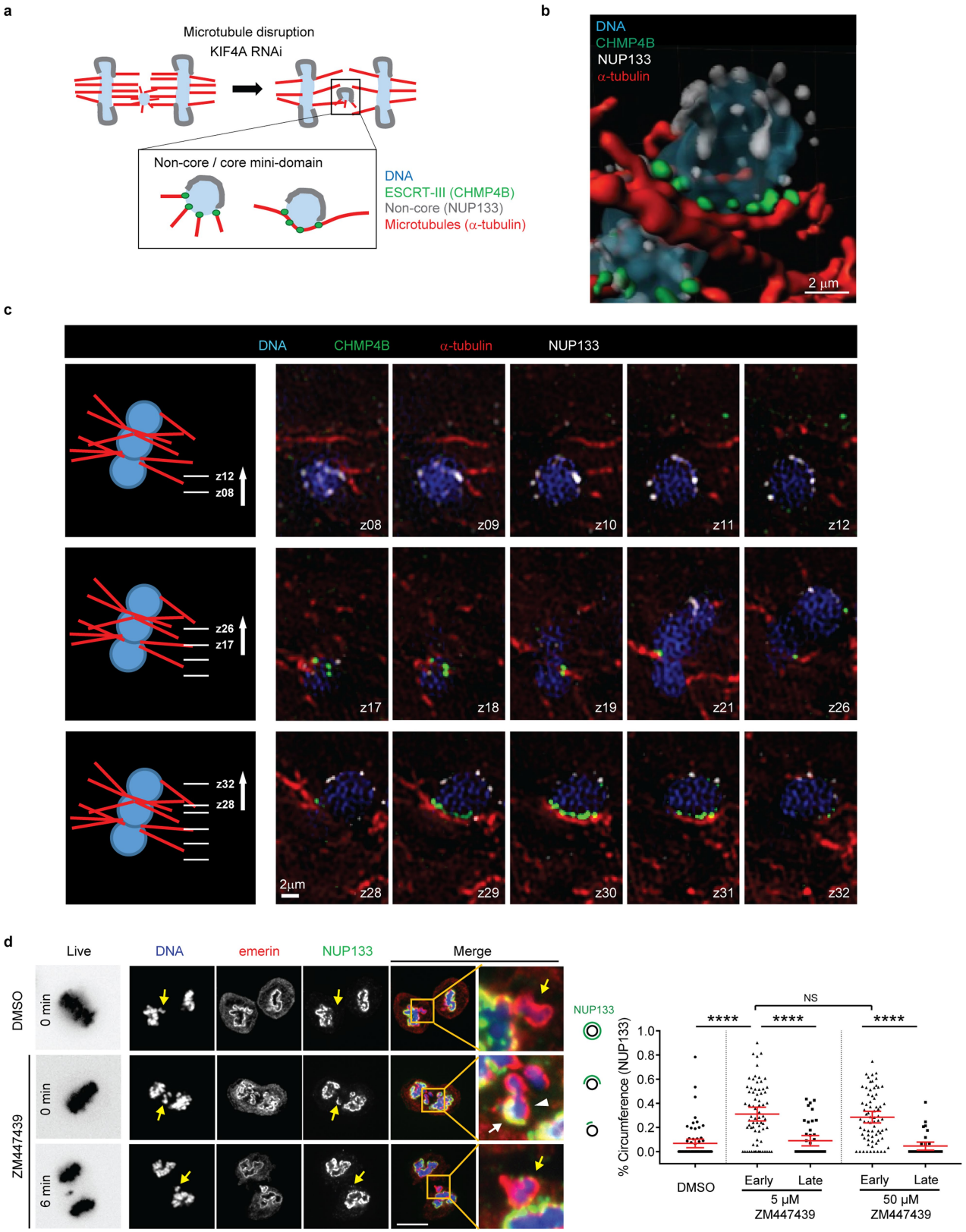


Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 7 | KIF4A knockdown partially restores non-core NE assembly to lagging chromosomes even when they have maximal (metaphase) levels of H3S10 phosphorylation. a, As reported[26], KIF4A knockdown largely preserves the redistribution of aurora B from centromeres to the central spindle (arrowheads) early in anaphase. Synchronization as in Fig. 1a. Left: Images from time-lapse series are shown from control (representative of 29 cells, from two experiments) or KIF4A-depleted (representative of 27 cells, from three experiments) HeLa K cells expressing aurora B–GFP (green) and RFP–H2B (red). $t = 0$ min, anaphase onset. Right, western blots showing depletion of KIF4A protein by siRNA in RPE-1 (top, three experiments) and HeLa K (bottom, three experiments) cells (also related to Fig. 3b, c; for gel source data see Supplementary Fig. 1). Scale bar, 10 μm. **b**, Restoration of LBR to some lagging chromosomes after KIF4A depletion (RPE-1 cells). Synchronization as in Fig. 1a. Left, representative images: DNA (blue), α-tubulin (red) and LBR (green). Right, MN/PN LBR FI ratio (mean with 95% CI, $n = 105, 119$, from three experiments). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. Scale bars, 10 μm. The restoration of LBR is often continuous around MN with partially restored MN/PN FI ratio (see also Fig. 2a after aurora B inhibition), whereas the NPC assembly is commonly restored discontinuously after KIF4A depletion (Fig. 3b, c) or aurora B inhibition (Extended Data Fig. 8d). This is consistent with different localization patterns of LBR and NPC in the core domain on the main chromosome mass (Extended Data Fig. 1b, c, e), which may arise from their different mobilities within the NE. **c**, Live-cell imaging confirming the partial restoration of an NPC protein (NUP107) to lagging chromosomes in HeLa K cells expressing GFP–NUP107 (green) and RFP–H2B (red) after KIF4A knockdown. Synchronization as in Fig. 1a. Images from a confocal time-lapse series; arrowheads, GFP–NUP107 protein recruited to lagging chromosomes (representative of six cells, from three experiments). Time shown in min. Scale bar, 10 μm. **d**, Unlike KIF4A knockdown, disruption of aurora B spindle midzone localization by MKPL2 knockdown fails to restore the recruitment of NUP133 to lagging chromosomes. Synchronization as in Fig. 1a. Graph shows quantification (mean with 95% CI, $n = 131, 136, 85$, from two experiments). $^{****}P < 0.0001$; NS, $P = 0.0616$, two-tailed Mann–Whitney test. **e**, H3S10 phosphorylation does not block NPC and non-core NE assembly. Left, representative images of KIF4A-depleted HeLa K cells showing restoration of non-core (NUP107) assembly onto lagging chromosomes with high level of H3S10 phosphorylation (red arrowhead, enlarged images of the blue boxed region). The level of phospho-H3S10 is similar between the lagging chromosomes with (red arrowhead) or without (white arrowhead) NUP107 recruitment. To illustrate the relative difference in pH3S10 levels between PN and MN, pH3S10 has been scaled differently in the pH3S10 channel and the merged channel. Cells were synchronized as in Extended Data Fig. 1e. Scale bar, 10 μm. Middle, linescan profile of indicated proteins along the dashed line in the merged image. Right, FI of phospho-H3S10 on MN and the corresponding PN (mean with 95% CI, $n = 24, 25, 25, 25, 24$, from two technical replicates). In MN that have recruited NUP107 (MN/PN FI ratio >0.4), the H3S10 phosphorylation level is comparable to metaphase chromosomes. $^{***}P = 0.0009$, $^{****}P < 0.0001$; NS, $P = 0.0709$, $P = 0.6699$, left to right; two-tailed Mann–Whitney test.

© 2018 Springer Nature Limited. All rights reserved.

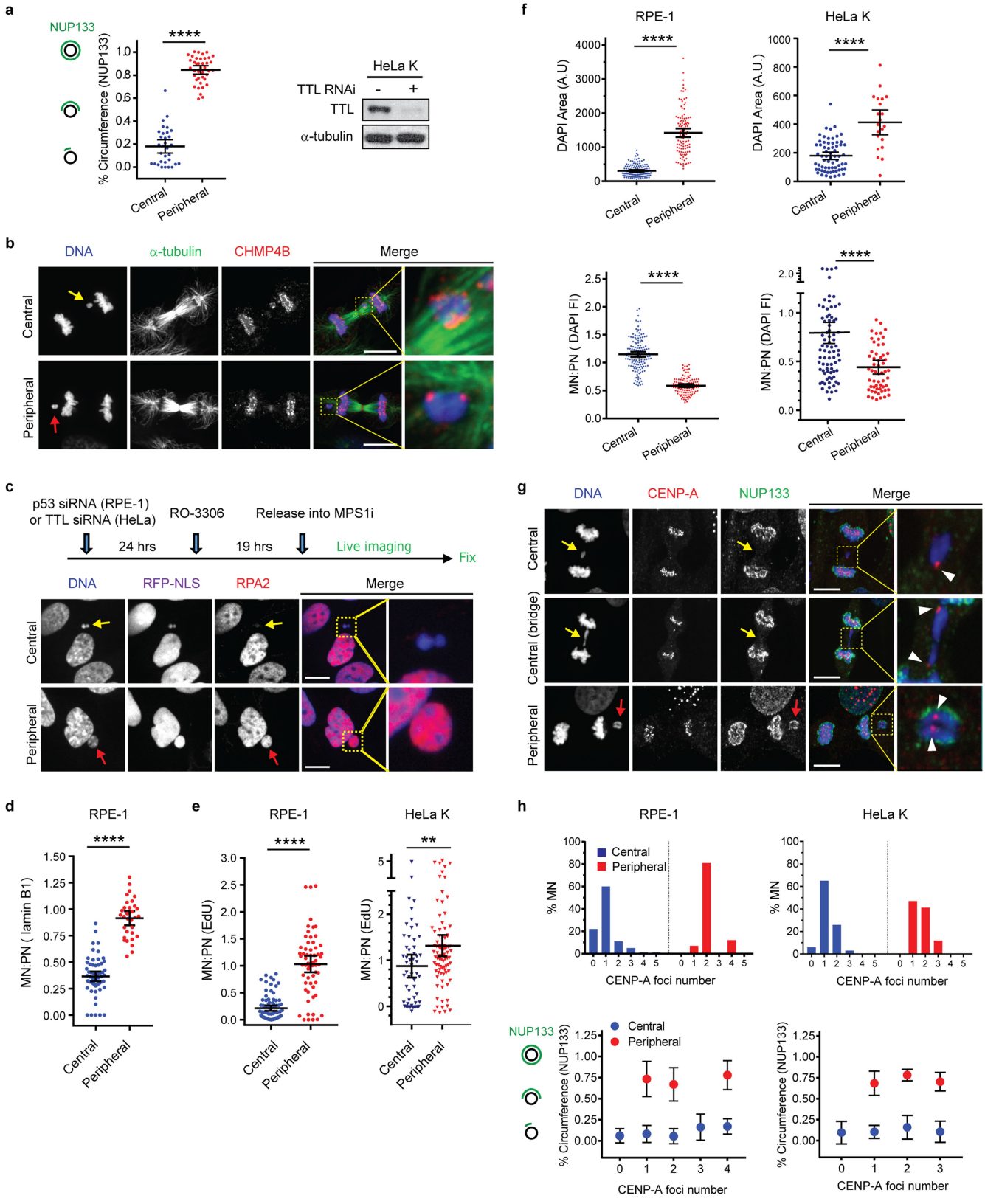


Extended Data Fig. 8 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 8 | Partitioning of the NE into core and non-core subdomains depends on the local organization of spindle microtubules. **a**–**c**, 3D-SIM showing exclusion of non-core NE assembly from the region of the lagging chromosome adjacent to CHIMP4B-marked microtubules. **a**, Summary of SIM results after KIF4A knockdown. The ESCRT-III complex is recruited to small membrane holes at the sites where spindle microtubules intersect the reforming NE, where it is thought to be required for normal NE sealing[23]. We used ESCRT-III (CHMP4B) labelling to identify the NE region on the lagging chromosome that is intersected by microtubules for the SIM experiments. **b**, Imaris surface 3D renderings from SIM images (from **c**) showing recruitment of NUP133 (white) to the region of a lagging chromosome depleted of CHMP4B (green) and microtubules (red) from a KIF4A-depleted HeLa K cell (representative of four lagging chromosomes, from two experiments). **c**, SIM images showing that, similar to main chromosome masses during normal NE assembly[23], NPCs (NUP133) are recruited to regions of the lagging chromosome lacking CHMP4B-decorated microtubules. Left, cartoon of entire lagging chromosome visualized from a KIF4A-depleted anaphase/telophase cell (microtubules in red, DNA in blue, *z* focal planes as indicated). Right, serial sections from a SIM *z*-stack of an anaphase/telophase HeLa K cell expressing CHMP4B–GFP (green) labelled for microtubules (red), NUP133 (white) and DNA (blue) after KIF4A depletion. These results are consistent with the idea that microtubules inhibit non-core NE assembly. **d**, KIF4A knockdown (Fig. 3b, c, Extended Data Fig. 7b–d) and early anaphase inhibition of aurora B have a similar effect on NPC and non-core NE assembly to lagging chromosomes. Experiments performed as in Fig. 2a in RPE-1 cells. Micrographs: left column, representative timelapse images of cells upon addition of ZM447439 or DMSO (live, 0 or 6 min after anaphase onset); right five columns, representative images of the same cells after fixation and labelling (12 min after treatment). For the cell exposed to ZM447439 at anaphase onset (middle row), one of the two lagging chromosomes exhibits small-scale separation of core (white arrowhead, emerin) and non-core (white arrow, NUP133) NE domains. Scale bar, 10 μm. Right, quantification (mean with 95% CI, $n = 64, 72, 43, 72, 34$, from two experiments). $^{****}P < 0.0001$; NS, $P = 0.6932$, two-tailed Mann–Whitney test.

© 2018 Springer Nature Limited. All rights reserved.

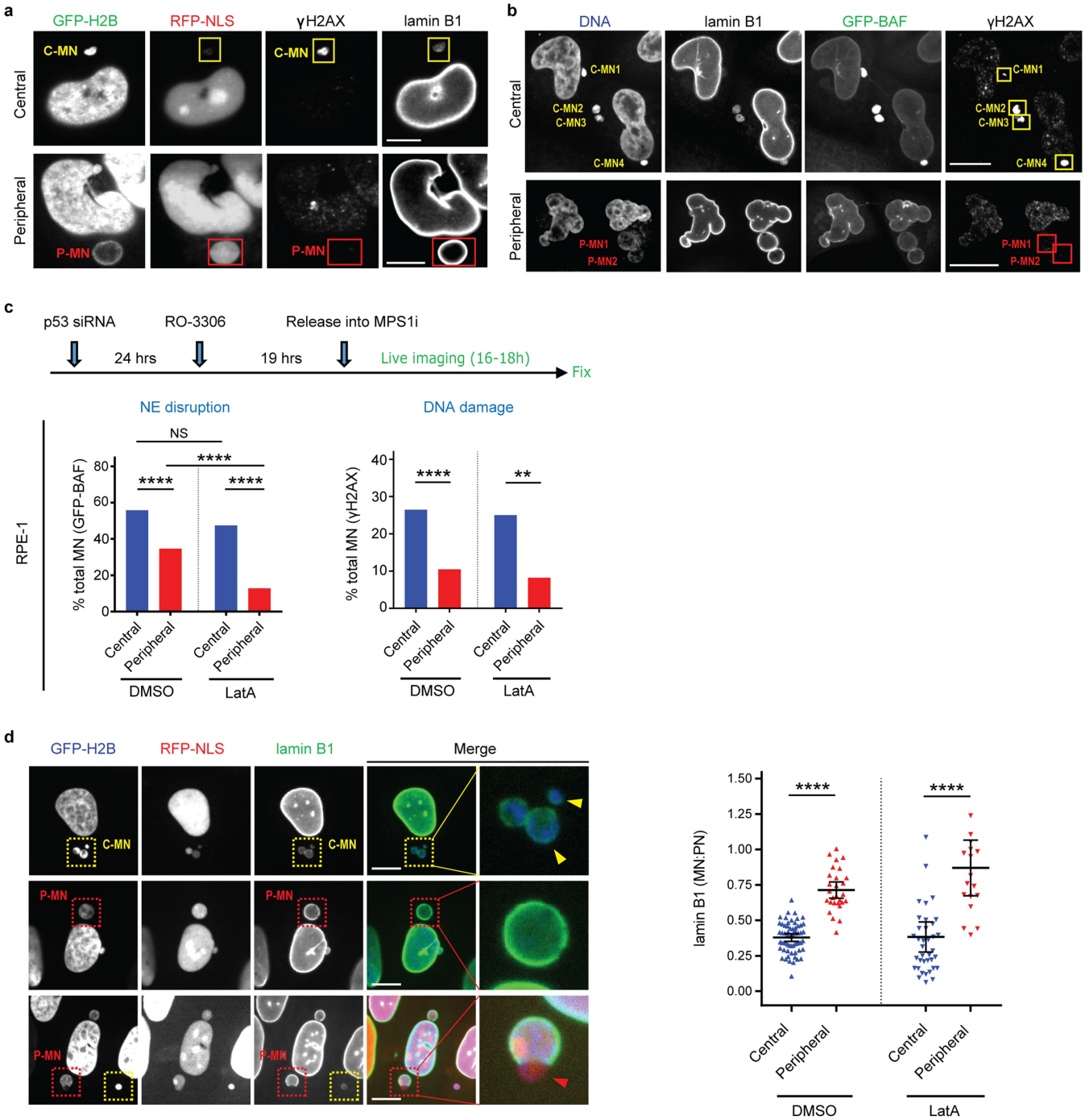


Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | Characterization of peripherally localized missegregated chromosomes and the resulting MN: microtubule proximity, nuclear import, and chromosome number. a, Restored assembly of NUP133 on MN derived from peripheral chromosomes as compared to central spindle lagging chromosomes in HeLa K cells. Left, quantification of the fraction of the lagging chromosome circumference with NUP133 (mean with 95% CI, $n = 32, 40$, from two experiments). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. Right, western blots showing depletion of TTL by siRNA in HeLa K cells (three experiments; for gel source data see Supplementary Fig. 1). **b**, Fewer microtubules associated with peripheral missegregated chromosomes (red arrow) than with MN from central missegregated chromosomes (yellow arrow). Synchronization as in Fig. 4a. Representative images of RPE-1 cells after glutaraldehyde fixation (representative of 30 central or peripheral missegregated chromosome, from two technical replicates). Scale bars, 10 μm. The sparse microtubule density near the peripheral chromosome (bottom images) was never associated with more than three CHMP4B foci in 30 images. **c**–**e**, Normal nuclear import and DNA replication in MN from peripherally missegregated chromosomes (red arrows) as compared to MN from central missegregated chromosomes (yellow arrows). **c**, Top, scheme of experiment. RPE-1 cells were treated with p53 siRNA to facilitate cell-cycle progression. After mitotic exit, micronucleated cells were imaged at 20-min intervals. Bottom, images of fixed and labelled RPE-1 cells 4 h after RO release at the end of the live-cell experiment. Scale bars, 10 μm. See Fig. 4d for quantification. **d**, MN/PN FI ratio for lamin B1 in RPE-1 cells, 4 h after RO release (mean with 95% CI, $n = 60, 32$, from two experiments). $^{****}P < 0.0001$, two-tailed Mann–Whitney test. **e**, Restoration of DNA replication (EdU incorporation, shown as the MN/PN FI ratio) in MN from peripheral chromosomes in RPE-1 (left) and HeLa K (right) cells, 16–18 h after RO release (mean with 95% CI, $n = 81, 60$, from five experiments for RPE-1; $n = 55, 86$, from two experiments for HeLa K). In RPE-1 cells, as reported[46], high dose MPS1 inhibition causes a fraction of cells to delay cell cycle progression. Therefore, we restricted analysis to cells that progressed into the S-phase of the cell cycle as evidenced by EdU labelling of the PN. $^{**}P < 0.0067$, $^{****}P < 0.0001$, two-tailed Mann–Whitney test. **f**, MN derived from peripheral chromosomes were larger (DAPI area, top) and more decondensed (density of DAPI FI, bottom) than MN from central chromosomes. Data are from RPE-1 (left) and HeLa K (right) cells (mean with 95% CI, $n = 145, 112$, from three experiments for DAPI area for RPE-1, $n = 145, 112$, from three experiments for DAPI FI for RPE-1; $n = 66, 21$, from two experiments for DAPI area for HeLa K, $n = 88, 58$, from three experiments for DAPI FI for HeLa K). $^{****}P < 0.0001$, two-tailed Welch's *t*-test for DAPI FI, two-tailed Mann–Whitney test for DAPI area. The extensive decondensation of MN from peripheral missegregated chromosomes may be because they initiate nuclear import slightly earlier than the assembling main daughter nuclei (Fig. 4c) and differ in surface area to volume ratio. **g**, **h**, Micronuclear chromosome number does not influence non-core NE assembly. Comparison of chromosome number in MN from central chromosomes and peripheral chromosomes. **g**, Confocal images of RPE-1 cells labelled for CENP-A (centromere marker, arrowheads), NUP133 and DNA. Yellow arrows, central chromosomes; red arrows, peripheral chromosomes. Synchronization as in Fig. 4a. Scale bars, 10 μm. Quantified in **h**. **h**, Top, distribution of chromosome number (CENP-A foci) in central (blue) and peripheral (red) MN. Data from RPE-1 (left) and HeLa K (right) cells. Bottom, fraction of lagging chromosome circumference with NUP133 for MN with the indicated chromosome number in RPE-1 (left) and HeLa K (right) cells (mean ± s.d., $n = 40, 111, 20, 9, 3$ each category for central chromosomes, $n = 0, 6, 71, 0, 11$ for peripheral chromosomes, from two experiments for RPE-1; mean ± s.d., $n = 2, 35, 14, 2, 0$ each category for central chromosomes, $n = 0, 8, 6, 2, 0$ for peripheral chromosomes, from two experiments for HeLa K). Although MN from peripheral chromosomes generally contain more chromosomes than MN from central chromosomes, non-core assembly is inhibited on MN from central chromosomes irrespective of their chromosome number. Furthermore, non-core NE assembly is restored to near-normal levels on MN from peripheral chromosomes, again irrespective of chromosome number.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | Peripheral localization of lagging chromosomes reduces MN NE disruption and DNA damage; actin-dependent peripheral MN disruption in RPE-1 cells. a, Peripheral localization of lagging chromosomes reduces MN disruption and DNA damage. Representative images of HeLa K cells (C-MN from central chromosomes; P-MN from peripheral chromosomes). See Fig. 4e for quantification. Scale bars, 5 μm. **b**, Detection of MN disruption by GFP–BAF accumulation yields similar results as in **a** for NE disruption detection by the loss of RFP–NLS. Representative images from a live-cell/fixed cell imaging experiment (as in Fig. 4e) from HeLa K cells expressing GFP–BAF. C-MN show NE disruption detected by hyper-accumulation of GFP–BAF (top) whereas P-MN show infrequent disruption (bottom). NE disruption of C-MN is accompanied by the acquisition of DNA damage (γH2AX, see Fig. 4e). At the end of 16–18 h live imaging, cells were fixed and stained for γH2AX and lamin B1. See Fig. 4e for quantification. Scale bars, 5 μm. **c**, In RPE-1 cells, P-MN undergo actin-dependent NE disruption. Top, experimental scheme. Chromosome missegregation was induced in RPE-1 cells expressing RFP–H2B and GFP–BAF (or GFP–H2B and RFP–NLS). About 1 h after mitotic exit, cells were treated with DMSO or a low dose (150 nM) of the actin assembly inhibitor latrunculin A (LatA) and imaged for 16–18 h. Bottom, percentage of P-MN or C-MN that underwent NE disruption (hyperaccumulation of GFP–BAF, from four experiments for DMSO and LatA) or that displayed DNA damage (FI of γH2AX in MN >3 s.d. of background γH2AX in PN, from six experiments for DMSO and three experiments for LatA). NS, $P = 0.0948$, $^{**}P = 0.0044$, $^{****}P < 0.0001$, two-tailed Fisher's exact test. For NE disruption: $n = 258$ for C- MN (DMSO), $n = 176$ for P-MN (DMSO), $n = 173$ for C-MN (LatA), $n = 125$ for P-MN (LatA); for DNA damage: $n = 306$ for C-MN (DMSO), $n = 182$ for P-MN (DMSO), $n = 128$ for C-MN (LatA), $n = 73$ for P-MN (LatA). **d**, P-MN develop discontinuities in the lamin B1 nuclear rim. Left, representative images of RPE-1 cells ~18 h after mitotic exit, as in **c**. Yellow boxes indicate C-MN (yellow arrowheads) with reduced lamin B1. Right, quantification (mean with 95% CI, $n = 61$, 28, 38, 18, left to right, from two experiments). $^{****}P < 0.0001$, two-tailed Welch's *t*-test for DMSO, two-tailed Mann–Whitney test for LatA. Red boxes indicate P-MN, one of which displays a prominent lamin B1 rim discontinuity. Red arrowhead in the enlarged image shows an NE herniation on a P-MN. Scale bars, 10 μm. Consistent with prior work, MN from lagging chromosomes undergo spontaneous disruption independent of actin[13]. By contrast, transient disruption of PN, as occurs during confined cell migration[47], is mediated by actomyosin contractile forces[13]. We confirm that MN from central lagging chromosomes undergo disruption independent from actin. However, we noted a difference between RPE-1 and HeLa K cells in the behaviour of MN from peripheral chromosomes. Although NE assembly appeared to be restored in both HeLa K and RPE-1 cells, P-MN in RPE-1 cells underwent residual disruption (there is nevertheless a statistically significant reduction of NE disruption frequency when comparing P-MN and C-MN in RPE-1 cells). We hypothesized that in highly motile RPE-1 cells, large P-MN might be more likely to undergo actin-dependent NE breakage, essentially becoming more similar to the transient NE disruption of PN[13,47]. The data in **c** confirm that P-MN undergo actin-dependent breakage. Furthermore, P-MN in RPE-1 cells have lamin B1 gaps (**d**, red arrowhead)[6]. One mechanism that could generate these gaps is residual contact between peripheral missegregated chromosomes and astral microtubules (Extended Data Fig. 9b). In addition, P-MN are more decondensed, and have larger NE surface area, than PN, which may dilute lamins. The increased breakage of P-MN in RPE-1 cells relative to HeLa K cells may be due to higher contractile forces in RPE-1 cells.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0538-8

mRNA circularization by METTL3–eIF3h enhances translation and promotes oncogenesis

Junho Choe[1,2,17], Shuibin Lin[1,2,3,17], Wencai Zhang[4], Qi Liu[1,2], Longfei Wang[2], Julia Ramirez-Moya[1,5], Peng Du[1,2], Wantae Kim[6,16], Shaojun Tang[7,8], Piotr Sliz[2], Pilar Santisteban[5], Rani E. George[9,10], William G. Richards[11], Kwok-Kin Wong[12], Nicolas Locker[13], Frank J. Slack[4,14,15] & Richard I. Gregory[1,2,10,14,15]*

N^6-methyladenosine (m^6A) modification of mRNA is emerging as an important regulator of gene expression that affects different developmental and biological processes, and altered m^6A homeostasis is linked to cancer[1–5]. m^6A modification is catalysed by METTL3 and enriched in the 3′ untranslated region of a large subset of mRNAs at sites close to the stop codon[5]. METTL3 can promote translation but the mechanism and relevance of this process remain unknown[1]. Here we show that METTL3 enhances translation only when tethered to reporter mRNA at sites close to the stop codon, supporting a mechanism of mRNA looping for ribosome recycling and translational control. Electron microscopy reveals the topology of individual polyribosomes with single METTL3 foci in close proximity to 5′ cap-binding proteins. We identify a direct physical and functional interaction between METTL3 and the eukaryotic translation initiation factor 3 subunit h (eIF3h). METTL3 promotes translation of a large subset of oncogenic mRNAs—including bromodomain-containing protein 4—that is also m^6A-modified in human primary lung tumours. The METTL3–eIF3h interaction is required for enhanced translation, formation of densely packed polyribosomes and oncogenic transformation. METTL3 depletion inhibits tumorigenicity and sensitizes lung cancer cells to BRD4 inhibition. These findings uncover a mechanism of translation control that is based on mRNA looping and identify METTL3–eIF3h as a potential therapeutic target for patients with cancer.

Despite awareness of the biological importance of m^6A in various organisms, the mechanisms by which m^6A regulates gene expression remain poorly understood. Proteins that contain YTH domains can specifically bind m^6A-modified RNA to regulate mRNA splicing, export, stability and translation[5–8]. In addition, we found that tethering METTL3 to mRNA reporters promotes translation and METTL3 enhances translation of oncogenes including *EGFR* in human lung cancer cells[1].

The initiation step of translation is rate limiting and a closed-loop model was proposed to facilitate multiple rounds of mRNA translation. This is supported by functional and physical interaction between the capped 5′ terminus and the polyadenylated 3′ terminus of mRNA mediated by the translation initiation factor eIF4G and PABPC1 (poly(A) binding protein cytoplasmic 1)[9–12]. Our data support an alternative closed-loop model, in which circularization of the mRNA is mediated by association between the eIF3h subunit at the 5′-end of the mRNA and METTL3 bound to specific sites near the translation stop codon. We find that METTL3 promotes translation of a large subset of oncogenic mRNAs and that METTL3–eIF3h interaction is required for oncogenic transformation. These findings provide new insights into the mechanism of translation control and suggest that METTL3–eIF3h and downstream oncogenes could be therapeutic targets for cancer.

To test a mRNA-looping model we first examined whether the METTL3-binding position on mRNA is important for its ability to enhance translation. Using reporters with MS2-binding sites located at different positions (Extended Data Fig. 1a), we found that direct METTL3 tethering can promote translation only when bound to the 3′ untranslated region (UTR) at a position near the stop codon. In vitro translation assays performed with cell lysates, recombinant METTL3 protein and in vitro-transcribed mRNAs confirmed that full-length METTL3 or an amino-terminal fragment (1–200 amino acids (a.a.)) that is sufficient to promote reporter translation in cells (Extended Data Fig. 2) enhanced translation of a luciferase reporter mRNA (Extended Data Fig. 1e–g). METTL3 tethering had a stronger effect on mRNAs that lack a poly(A) tail (Extended Data Fig. 1g), which is consistent with some redundancy between METTL3 and eIF4G–PABPC1-mediated looping for mRNAs with poly(A) tails in these in vitro assays.

To provide evidence that METTL3 is bound to polyribosomes, mRNA ribonucleoprotein complexes (mRNPs) containing Flag–METTL3 were affinity purified, incubated with gold-labelled anti-METTL3 antibodies, subjected to sucrose gradient fractionation and then analysed by electron microscopy (Fig. 1a). This revealed gold-labelled METTL3 in the individual polyribosomes (Extended Data Fig. 4a, b). We performed similar experiments using either anti-CBP80 or anti-eIF4E gold-labelled antibodies together with the anti-METTL3 particles. As the anti-CBP80 and anti-eIF4E gold particles were larger, they could be distinguished from the anti-METTL3 particles. Individual polyribosomes containing double-labelled gold particles showed that each METTL3 signal is in close proximity (less than 20 nm) to a cap-binding protein (Fig. 1b, Extended Data Fig. 4c, d). This reveals the topology of individual endogenous METTL3-bound polyribosomes and supports the notion that METTL3 mediates the looping of mRNA to promote efficient translation.

Full-length METTL3 as well as the 1–200 a.a. and 1–350 a.a. fragments were found to associate with m^7GTP-agarose in cap-binding assays (Extended Data Fig. 3a). This result is highly consistent with tethering assays (Extended Data Fig. 2) and supports the notion that the 1–200 a.a. fragment of METTL3 interacts with translation initiation factor(s) to promote translation. Knockdown of METTL3 had no effect on the association of cap-binding proteins or translation initiation factors (Extended Data Fig. 3b). Thus, formation of the translation initiation complex does not require METTL3. Conversely, the association of METTL3 with m^7GTP-agarose was diminished using lysates depleted for CTIF, eIF4GI or eIF3b, supporting the notion that the

[1]Stem Cell Program, Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. [3]Center for Translational Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. [4]Department of Pathology, Cancer Center, Beth Israel Deaconess Medical Center, Boston, MA, USA. [5]Instituto de Investigaciones Biomédicas, Consejo Superior de Investigaciones Científicas and Universidad Autónoma de Madrid (CSIC-UAM), Madrid, Spain. [6]Harvard School of Dental Medicine, Boston, MA, USA. [7]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA. [8]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA. [9]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [10]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [11]Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA. [12]Division of Hematology and Medical Oncology, NYU School of Medicine, New York, NY, USA. [13]School of Biosciences and Medicine, University of Surrey, Guildford, UK. [14]Harvard Initiative for RNA Medicine, Boston, MA, USA. [15]Harvard Stem Cell Institute, Cambridge, MA, USA. [16]Present address: Biomedical Translational Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, South Korea. [17]These authors contributed equally: Junho Choe, Shuibin Lin. *e-mail: rgregory@enders.tch.harvard.edu

© 2018 Springer Nature Limited. All rights reserved.

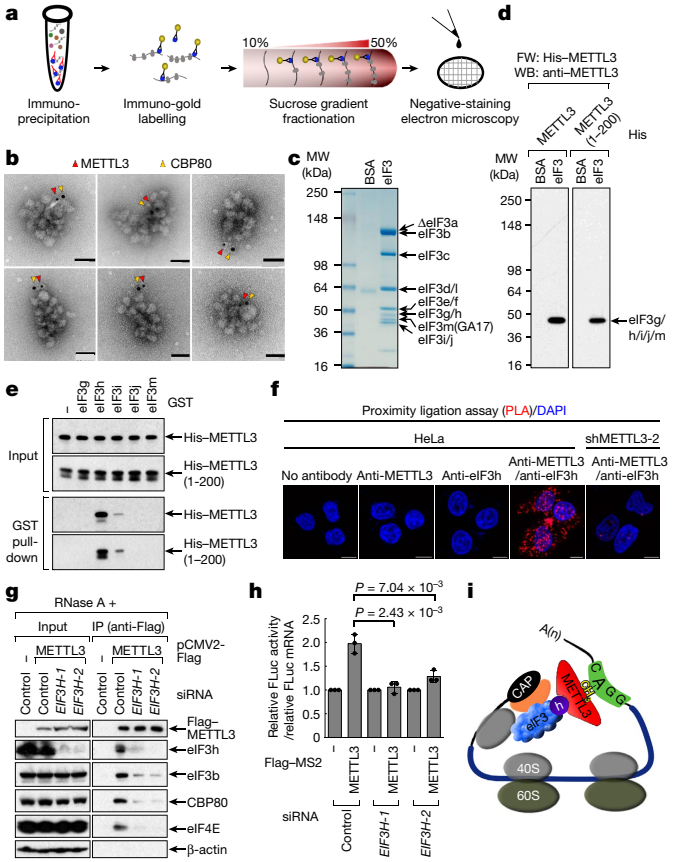


Fig. 1 | METTL3 enhances translation of target mRNAs by interacting with eIF3h. a, Electron microscopy procedure. **b**, Electron-microscopy images of polyribosomes. Red arrows, METTL3 with immuno-gold particle (6 nm); yellow arrows, CBP80 with immuno-gold particle (10 nm). Scale bars, 50 nm. Images are representative of three independently performed experiments with similar results. **c**, **d**, Far-western (FW) blot. **c**, Colloidal Coomassie blue staining of eIF3 complex. A breakdown product is denoted (ΔeIF3a). **d**, Far-western blot of purified eIF3 complex. Images are representative of two independently performed experiments with similar results. **e**, GST-tagged eIF3 subunits and co-purified His–METTL3 or 1–200 a.a. analysed by western blotting. Two independently performed experiments show similar results. **f**, Proximity ligation assay. Images are representative of two independently performed experiments with similar results. shMETTL3-2, METTL3 knockdown mediated by shRNA against *METTL3*; -1 and -2 suffixes denote separate identical shRNAs. **g**, Co-immunoprecipitation from control or eIF3h knockdown cells. Images are representative of two independently performed experiments with similar results. IP, immunoprecipitate. **h**, Tethering assays. Data are mean ± s.d. from three biologically independent samples. *P* values are from a two-sided *t*-test. FLuc, firefly luciferase. **i**, Model of METTL3-dependent mRNA translation. CAP, cap; h, eIF3h; GGAC, m6A motif sequence; (A)n, poly(A) tail; siRNA, small interfering RNA; 40S, 40S ribosomal subunit; 60S, 60S ribosomal subunit.

association of METTL3 with m^{7}GTP-agarose is mediated through an interaction with general translation initiation factor(s) (Extended Data Fig. 3c).

A large-scale purification and mass spectroscopy characterization of complexes containing Flag–METTL3 identified numerous translation factors (Extended Data Fig. 3d). Gene Ontology analysis of the METTL3-interacting proteins identified 'mRNA metabolic processes', 'RNA processing' and 'Translation' as the most significantly enriched categories (Extended Data Fig. 3e, f). Considering this and our previous observation that METTL3 knockdown diminishes the association of eIF3 with cap-binding proteins in co-immunoprecipitation, we hypothesized that METTL3 might interact directly with certain component(s) of the multi-subunit eIF3 complex.

To test whether METTL3 interacts with any of the 13 subunit(s) of eIF3, recombinant METTL3 and 1–200 a.a. were used for far-western blotting with a purified human eIF3 complex (Extended Data Fig. 4e, Fig. 1c). METTL3 and 1–200 a.a. both specifically bound to a single band that probably corresponds to eIF3g, eIF3h, eIF3i, eIF3j or eIF3m (Fig. 1d). To further confirm this interaction and to define the particular subunit(s) that interacts with METTL3, we individually expressed and purified the GST-tagged eIF3 subunits from bacteria (Extended Data Fig. 4f) and tested them for binding to His–METTL3 using in vitro binding assays with either His–METTL3 or 1–200 a.a. METTL3 (and 1–200 a.a.) were found to specifically interact with eIF3h (Fig. 1e). The interaction between eIF3h with METTL3 was further confirmed using an anti-METTL3 antibody (that recognizes a 1–250 a.a. METTL3 epitope) to specifically disrupt this eIF3h–METTL3 interaction (Extended Data Fig. 4g). Additional GST pull-down experiments identified the MPN (Mpr1p and Pad1p N-terminal) domain[13] as necessary and sufficient to interact with METTL3 (Extended Data Fig. 4h–j). Notably, the MPN domain faces the solvent side of the ribosome and is probably accessible for interaction with METTL3 without impairing 80S assembly[13].

An in situ proximity ligation assay (PLA) confirmed METTL3–eIF3h proximity in cells (Fig. 1f). Co-immunoprecipitation experiments revealed that the association of METTL3 with translation initiation factors is dependent on eIF3h (Fig. 1g). We next tested the functional interaction between METTL3 and eIF3h, and found that depletion of eIF3h abrogated enhanced translation by METTL3 (Fig. 1h, Extended Data Fig. 4k–m). Overall, our results support a model whereby mRNA circularization and translation control is mediated by a specific METTL3–eIF3h interaction (Fig. 1i).

We next examined the widespread effect of METTL3 on mRNA translation (Fig. 2, Extended Data Fig. 5). METTL3 depletion caused an increase in the 80S ribosome peak and a corresponding reduction of polyribosome peak (Fig. 2a). METTL3 depletion had a negligible effect on steady-state mRNA abundance whereas translation efficiency of a large subset (4,267) of mRNAs was reduced by at least twofold in METTL3-depleted cells (Fig. 2b–d). Comparison of these genes with previously reported METTL3 photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) data[14] identified 809 mRNAs that are both bound and translationally regulated by METTL3 (Fig. 2d). mRNAs on this list of METTL3 targets have on average longer 3′ UTRs compared to all mRNAs (Fig. 2e). Gene Ontology showed that these mRNAs are involved in tumour progression and apoptosis (Extended Data Fig. 5b). Quantitative PCR with reverse transcription (RT–qPCR) analysis confirmed that METTL3 depletion had a modest effect on mRNA abundance (Extended Data Fig. 5c), but strongly decreased translation of target mRNAs (Fig. 2f). Moreover, neither global mRNA stability analysis by RNA sequencing (RNA-seq) (Fig. 2g) nor RT–qPCR analysis of individual genes (Extended Data Fig. 5d, e) showed any differences in mRNA stability upon METTL3 depletion. We isolated endogenous METTL3 mRNPs and confirmed the association with target mRNAs by RT–qPCR (Fig. 2h, Extended Data Fig. 6a). Western blotting showed decreased protein expression from these target mRNAs in the METTL3-knockdown samples (Fig. 2i), whereas analysis of mRNA sequencing data revealed that the splicing patterns of these target mRNAs was unaltered (Extended Data Fig. 6b). Depletion of YTHDF1, a m^{6}A-reader protein implicated in translational control, had no effect on the expression of these METTL3 targets[7] (Extended Data Fig. 6c). Western blotting showed strongly reduced endogenous expression of bromodomain-containing protein 4 (BRD4) upon knockdown of eIF3h without affecting *BRD4* mRNA abundance or the levels of METTL3 protein (Fig. 2j, Extended Data Fig. 6d). Expression of a short-hairpin RNA (shRNA)-resistant Flag–METTL3 rescued the expression of BRD4 and CD9 proteins. However, expression of a catalytic mutant of METTL3 failed to recover target gene expression (Fig. 2k). METTL3 knockdown in A549 lung cancer cells similarly led to decreased expression of BRD4 and other targets (Fig. 2l), and treatment of METTL3-depleted

© 2018 Springer Nature Limited. All rights reserved.

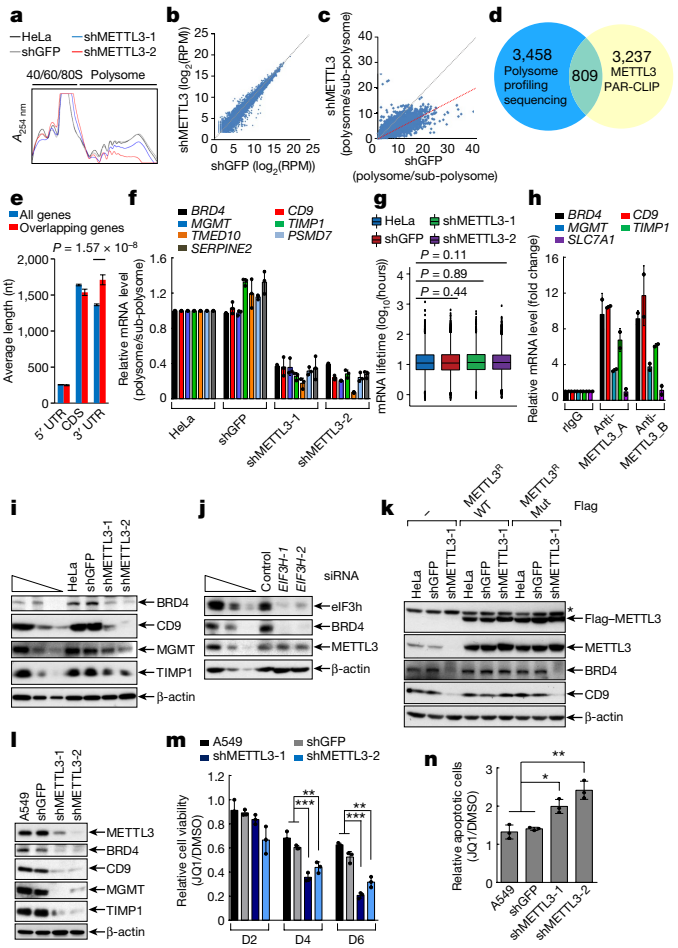


Fig. 2 | METTL3 promotes translation of a large subset of mRNAs. **a**, Polysome profile. Plot is representative of two independently performed experiments with similar results. **b**, Scatter plot of RNA-seq data. Average read number from two METTL3 knockdowns is plotted. As a control, shRNA targeting *GFP* (shGFP) was used. **c**, Scatter plot of translation efficiency. Average read number from two shMETTL3 samples were calculated by the ratio of polysome fraction to the read number in the sub-polysome fraction (40/60/80S) and plotted on the y-axis. **d**, Venn diagram showing mRNAs with a greater than twofold change in translation efficiency and with METTL3 PAR-CLIP data. **e**, Features of overlapping mRNAs ($n = 809$) from **d** was compared with all (18,115) expressed genes. Data are mean ± s.e.m. *P* values are from an unpaired two-sided *t*-test. **f**, RT–qPCR analysis. Data are mean ± s.d. from three technical replicates. **g**, Box plot represents stability profiling of global mRNA (12,479 mRNAs) from two biological replicates. *P* values are from a two-sided Wilcoxon rank sum test; the results are not statistically significant. **h**, RT–qPCR analysis of METTL3-associated mRNAs using two different anti-METTL3 antibodies (anti-METTL3_A and anti-METTL3_B). Data are mean ± s.d. from two independent experiments. **i–l**, Western blots. Cells were expressed with indicated shRNA, siRNA or plasmids. Blots are representative of at least two independently performed experiments with similar results. **m**, MTS assay of A549 cellular proliferation upon JQ1 treatment. **n**, Quantification of (sum of early and late) apoptotic cells. **m**, **n**, Data are mean ± s.d. from three independent experiments. *P* values are from a two-sided *t*-test with multiple comparison; $***P < 0.001$, $**P < 0.01$, $*P < 0.05$.

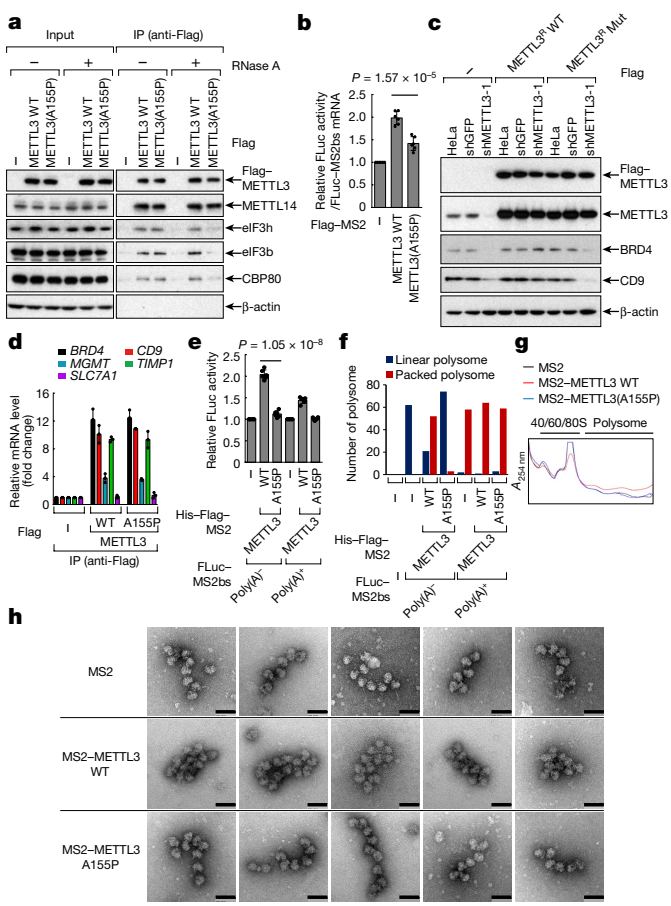


Fig. 3 | METTL3–eIF3h interaction is crucial for enhanced mRNA translation and polysome conformation. **a**, Co-immunoprecipitation of Flag–METTL3 or Flag–METTL3(A155P) analysed by western blotting using the indicated antibodies. Where indicated, lysates were treated with RNase A. Blots are representative of two independently performed experiments with similar results. **b**, Tethering assay to measure translation efficiency of reporter mRNAs. Data are mean ± s.d. from six independent experiments. *P* values are from a two-sided *t*-test. **c**, Western blot analysis using indicated antibodies. Two independently performed experiments show similar results. METTL3^R, cDNA containing nucleotide substitutions that make it resistant to shRNA for expression of wild-type or mutant METTL3 protein in the stable knockdown cells. **d**, RT–qPCR. Data are mean ± s.d. from two independent experiments. **e**, In vitro translation with rabbit reticulocyte lysate. Data are mean ± s.d. from six independent experiments. *P* values are from a two-sided *t*-test. **f**, Analysis of 20 images from each sample in **e**. **g**, Peak analysis of polysome profiling coupled with in vitro translation. Data are representative of two independently performed experiments with similar results. **h**, Electron microscopy images of polyribosomes. Images were taken from the samples in **g**. Scale bars, 50 nm. Images are representative of two independently performed experiments with similar results.

and control A549 cells with the BRD4 inhibitor JQ1 revealed that the METTL3-depleted cells are more sensitive to pharmacological BRD4 inhibition (Fig. 2m, n, Extended Data Fig. 6e).

Considering that 1–200 a.a. is sufficient to directly interact with eIF3h (Fig. 1 and Extended Data Fig. 4) and that 1–200 a.a. can promote translation in tethering experiments whereas 1–150 a.a. does not (Extended Data Fig. 2), we reasoned that a region between 150–200 a.a. must be important for the physical and functional METTL3–eIF3h interaction. Secondary structure predictions identified a putative alpha helix (150–161 a.a.) that is highly conserved in mammals[15,16] (Extended Data Fig. 7a, b). Moreover, 3D modelling identified a putative structured module[17] (Extended Data Fig. 7c). We therefore generated a mutant version of METTL3 with a single amino acid substitution of a highly conserved alanine (A155P) to disrupt this putative helical structure. Co-immunoprecipitation confirmed substantially impaired interaction of METTL3(A155P) with translation initiation factors, specifically in RNase-treated samples (Fig. 3a). Notably, the A155P mutation did not disrupt METTL3 interaction with METTL14 (Fig. 3a). Tethering experiments demonstrated that the METTL3(A155P) mutant is strongly impaired in promoting mRNA translation (Fig. 3b, Extended Data Fig. 7d, e). Moreover, expression of METTL3(A155P) in METTL3-depleted cells failed to rescue the expression of endogenous

© 2018 Springer Nature Limited. All rights reserved.

target proteins (Fig. 3c). This effect was not due to altered mRNA association of METTL3(A155P) with endogenous target mRNAs (Fig. 3d and Extended Data Fig. 7f). In summary, although the METTL3 A155P mutant can associate with METTL14 and be loaded on mRNAs, its ability to interact with initiation factors and promote mRNA translation is severely compromised.

We next used in vitro-reconstituted translation assays to explore the effect of 3′ UTR-bound METTL3 on translation efficiency and polysome conformation (Fig. 3e, f, Extended Data Figs. 7g, 9a). Consistent with Extended Data Fig. 1g, tethering of recombinant METTL3 increased translation efficiency with a stronger effect on mRNA lacking a poly(A) tail (Fig. 3e). By contrast, tethered METTL3(A155P) had no significant effect on translation (Fig. 3e). Electron microscopy analysis of samples from the in vitro translation reactions showed mostly densely packed polysome structures formed when METTL3 was tethered to reporter mRNAs lacking a poly(A) tail (Extended Data Fig. 9a). By contrast, almost all the polysomes observed in the METTL3(A155P) sample appeared as linear polysomes (as in the control sample). Only dispersed ribosomes were observed in reactions without mRNA (Fig. 3f, Extended Data Fig. 9a). In all samples with mRNAs with poly(A) tails, mostly packed polysomes were observed irrespective of the protein that was tethered. As the reporter mRNA contains a very short 3′ UTR, it is expected that the known eIF4GI–PABP interaction is probably responsible for observed packed polysomes in poly(A)$^+$ samples (Fig. 3f, Extended Data Fig. 9a). We further analysed in vitro translation assays by sucrose gradient fractionation. Tethering of wild-type METTL3 resulted in a larger polysome peak compared to the METTL3(A155P) mutant (Fig. 3g). Individual polysomes were analysed using electron microscopy (Fig. 3g, h). Polysomes formed with wild-type METTL3 appeared to be more densely packed compared to the more linear polysomes that formed with the A155P mutant or MS2 control (Fig. 3h). Taken together, these results strongly support our model that METTL3 through its interaction with eIF3h promotes translation through its effects on polysome conformation.

Considering that METTL3 regulates the translation of a large subset of genes that are involved in tumour progression and apoptosis (Extended Data Fig. 5b), we examined the role of METTL3 in cancer. Immunohistochemistry staining of primary human lung adenocarcinoma samples and adjacent normal control tissue revealed that METTL3 expression is significantly increased in lung tumours and correlates with tumour stage (Fig. 4a, b, Extended Data Fig. 8a). METTL3 depletion in A549 lung cancer cells resulted in significantly smaller tumours in mouse xenografts (Fig. 4c, Extended Data Fig. 8b–d).

We next examined the relevance of the METTL3–eIF3h interaction and mRNA looping in the context of cancer-cell biology. Knockdown of eIF3h suppressed the ability of METTL3 to promote cellular invasion (Fig. 4d, Extended Data Fig. 8e). Moreover, unlike wild-type METTL3 protein, expression of either a catalytically inactive mutant or the METTL3(A155P) mutant was unable to promote the invasive capability of lung fibroblasts (Fig. 4e, Extended Data Fig. 8f). Furthermore, METTL3 overexpression was sufficient to promote the oncogenic transformation of NIH-3T3 cells, mouse embryonic fibroblasts (MEFs) or MB352 (p53-null MEFs) cells, whereas METTL3(A155P) had no significant effect in these 3D soft agar colony formation assays (Fig. 4f–h, Extended Data Fig. 8g, h). The oncogenic function of METTL3 was also studied in the mouse xenografts. NIH-3T3 cells with ectopic expression of wild-type METTL3, METTL3(A155P) or the empty-vector control were injected into nude mice to determine their in vivo tumorigenic capacities. Overexpression of wild-type METTL3 promoted in vivo tumour growth, whereas METTL3(A155P) showed an impaired ability to promote tumour growth. As control, no tumours were detected in mice injected with control NIH-3T3 cells (Fig. 4i, Extended Data Fig. 8i). Overall, these results support that the METTL3–eIF3h-mediated mRNA looping is critical for the oncogenic function of METTL3.

Methylated-RNA immunoprecipitation with sequencing (meRIP–seq) performed on four primary human lung tumours identified patient-specific and commonly m^{6}A-modified mRNAs in lung cancer

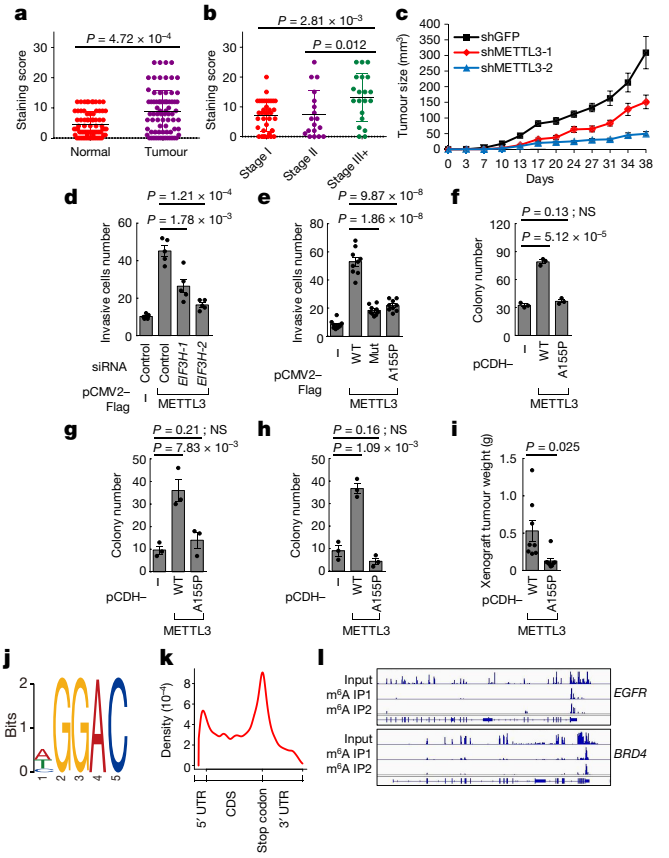


Fig. 4 | Role of METTL3 and m^{6}A in lung cancer cells and primary human tumours. a, **b**, Analysis of METTL3 immunohistochemical staining in primary lung adenocarcinoma and adjacent normal tissue. **a**, Normal tissue versus tumour tissue. Data are mean ± s.d. from 75 tissue samples. *P* value from a two-sided Wilcoxon signed-rank test. **b**, METTL3 levels in tumours of various stages. Data are mean ± s.d.; Stage I, $n = 37$; Stage II, $n = 18$; Stage III, $n = 20$. *P* values are from two-sided Wilcoxon signed-rank tests. **c**, Tumour growth of xenografts from A549 cells stably expressing indicated shRNAs. Data are mean ± s.e.m. from 5 mice. **d**, **e**, Quantification of invasive BJ cells. Cells were transiently transfected with indicated siRNAs (**d**) or plasmids (**e**). Data are mean ± s.e.m. from five (**d**) or nine (**e**) independent experiments. *P* values are from two-sided *t*-tests. **f**, Quantification of colony formation of NIH-3T3. **g**, **h**, Quantification of MEFs (**g**) or MB352 (**h**) colony formation. **f**–**h**, Data are mean ± s.e.m. from three independent experiments. *P* values from a two-sided *t*-test; NS, not significant. **i**, Tumour weight of xenografts derived from NIH-3T3 cells stably expressing the indicated proteins. There was no tumour formation in the empty vector group during the observed period. Data are mean ± s.e.m. from eight independent mice. *P* value is from a two-sided *t*-test. **j**–**l**, Global profiling of m^{6}A targets in primary lung cancer samples. **j**, Sequence motif identified in m^{6}A meRIP–seq. **k**, Metagene analysis of m^{6}A peaks. **l**, Integrative genomics viewer plots of representative m^{6}A-containing genes. Data is representative of four lung tumours with similar results.

(Extended Data Fig. 8j, Supplementary Table 3). An expected GGAC motif was identified and m^{6}A peaks are predominantly localized near the translation stop codon (Fig. 4j, k, Extended Data Fig. 8k). Gene Ontology analysis revealed that the common methylated genes are enriched in the signature genes of neoplasms and cancer, including the known oncogenes *EGFR* and *BRD4* (Fig. 4l, Extended Data Fig. 9b). These data were consistent with m^{6}A features identified in cell lines, with more than 50% of the peaks found in both tumours and cancer cell lines (Extended Data Fig. 9c).

We propose that METTL3 promotes oncogene translation and tumorigenesis through a mRNA looping mechanism. This is supported by: 1) the position-dependent effects of METTL3 tethering on mRNA translation; 2) electron-microscopy visualization of METTL3 bound

© 2018 Springer Nature Limited. All rights reserved.

to endogenous polyribosomes and its proximity to cap-binding proteins; 3) METTL3 interaction with eIF3h and 4) the disruption of METTL3–eIF3h interaction abolishes the ability of METTL3 to promote translation, affect polysome conformation or promote oncogenic transformation. This METTL3–eIF3h loop presumably promotes translation through ribosome recycling in a way similar to that proposed for eIF4G–PABPC1-mediated mRNA looping. Although there is probably some redundancy between these mRNA circularization mechanisms, looping between the stop codon and the 5′ end might represent a more productive way to recycle ribosomes rather than via the 3′ end, especially for mRNAs with long 3′ UTRs as the ribosomes will dissociate from the mRNP once released at the stop codon.

The complexity of eIF3 and the specialized roles of individual subunits are only beginning to be appreciated[18,19]. Imbalanced expression of eIF3 subunits is found in various tumours[20] and expression in fibroblasts can promote malignant transformation[21]. Increased eIF3h expression is found in different tumour types and in many cases is due to amplification of a chromosomal region that includes *EIF3H* at 8q23.3 and the nearby *MYC* oncogene[22,23]. METTL3 is overexpressed in many different types of cancer (Extended Data Fig. 10a), and the expression of METTL3 and eIF3h is often positively correlated (Extended Data Fig. 10b, c). Our findings could pave the way for the development of new cancer therapeutic strategies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0538-8.

Received: 7 July 2017; Accepted: 14 August 2018;
Published online 19 September 2018.

1. Lin, S., Choe, J., Du, P., Triboulet, R. & Gregory, R. I. The m^{6}A methyltransferase METTL3 promotes translation in human cancer cells. *Mol. Cell* **62**, 335–345 (2016).
2. Barbieri, I. et al. Promoter-bound METTL3 maintains myeloid leukaemia by m^{6}A-dependent translation control. *Nature* **552**, 126–131 (2017).
3. Vu, L. P. et al. The *N*^{6}-methyladenosine (m^{6}A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat. Med.* **23**, 1369–1376 (2017).
4. Chen, M. et al. RNA *N*^{6}-methyladenosine methyltransferase METTL3 promotes liver cancer progression through YTHDF2 dependent post-transcriptional silencing of SOCS2. *Hepatology* **67**, 2254–2270 (2017).
5. Meyer, K. D. & Jaffrey, S. R. The dynamic epitranscriptome: *N*^{6}-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* **15**, 313–326 (2014).
6. Wang, X. et al. *N*^{6}-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117–120 (2013).
7. Wang, X. et al. *N*^{6}-methyladenosine modulates messenger RNA translation efficiency. *Cell* **161**, 1388–1399 (2015).
8. Xu, C. et al. Structural basis for selective binding of m^{6}A RNA by the YTHDC1 YTH domain. *Nat. Chem. Biol.* **10**, 927–929 (2014).
9. Imataka, H., Gradi, A. & Sonenberg, N. A newly identified N-terminal amino acid sequence of human eIF4G binds poly(A)-binding protein and functions in poly(A)-dependent translation. *EMBO J.* **17**, 7480–7489 (1998).
10. Wells, S. E., Hillner, P. E., Vale, R. D. & Sachs, A. B. Circularization of mRNA by eukaryotic translation initiation factors. *Mol. Cell* **2**, 135–140 (1998).
11. Borman, A. M., Michel, Y. M., Malnou, C. E. & Kean, K. M. Free poly(A) stimulates capped mRNA translation *in vitro* through the eIF4G-poly(A)-binding protein interaction. *J. Biol. Chem.* **277**, 36818–36824 (2002).
12. Tarun, S. Z., Jr. & Sachs, A. B. Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. *EMBO J.* **15**, 7168–7177 (1996).
13. des Georges, A. et al. Structure of mammalian eIF3 in the context of the 43S preinitiation complex. *Nature* **525**, 491–495 (2015).
14. Liu, J. et al. A METTL3–METTL14 complex mediates mammalian nuclear RNA *N*^{6}-adenosine methylation. *Nat. Chem. Biol.* **10**, 93–95 (2014).
15. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
16. Kouza, M., Faraggi, E., Kolinski, A. & Kloczkowski, A. The GOR method of protein secondary structure prediction and its application as a protein aggregation prediction tool. *Methods Mol. Biol.* **1484**, 7–24 (2017).
17. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).
18. Lee, A. S., Kranzusch, P. J. & Cate, J. H. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature* **522**, 111–114 (2015).
19. Lee, A. S., Kranzusch, P. J., Doudna, J. A. & Cate, J. H. eIF3d is an mRNA cap-binding protein that is required for specialized translation initiation. *Nature* **536**, 96–99 (2016).
20. Hershey, J. W. The role of eIF3 and its individual subunits in cancer. *Biochim. Biophys. Acta* **1849**, 792–800 (2015).
21. Zhang, L., Pan, X. & Hershey, J. W. Individual overexpression of five subunits of human translation initiation factor eIF3 promotes malignant transformation of immortal fibroblast cells. *J. Biol. Chem.* **282**, 5790–5800 (2007).
22. Mahmood, S. F. et al. A siRNA screen identifies *RAD21*, *EIF3H*, *CHRAC1* and *TANC2* as driver genes within the 8q23, 8q24.3 and 17q23 amplicons in breast cancer with effects on cell growth, survival and transformation. *Carcinogenesis* **35**, 670–682 (2014).
23. Zhang, L., Smit-McBride, Z., Pan, X., Rheinhardt, J. & Hershey, J. W. An oncogenic role for the phosphorylated h-subunit of human translation initiation factor eIF3. *J. Biol. Chem.* **283**, 24047–24060 (2008).

Acknowledgements Thanks to Y. K. Kim for CBP80 and CTIF antibodies. S.L. was supported by a Damon Runyon-Sohn Pediatric Fellowship (DRSG-7–13) and a grant from Alex's Lemonade Stand Foundation. R.I.G. was supported by grants from the US National Institute of General Medical Sciences (NIGMS) (R01GM086386) and National Cancer Institute (NCI) (R01CA211328).

Reviewer information *Nature* thanks C. Mason and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.C., S.L. and R.I.G. designed the research; J.C., S.L., W.Z., L.W., J.R.-M. and W.K. performed all the experiments; J.C. performed in vitro translation assay, tethering assay, polysome fractionation and RNA-seq, co-immunoprecipitation, electron microscopy, cap-association assay and in situ PLA; S.L. performed GST pull-down assay, far-western blotting, immunohistochemical staining, cell invasion assay and m^{6}A meRIP–seq; W.Z. performed soft agar colony formation assays and in vivo tumour xenograft; L.W. performed electron microscopy; J.R.-M. performed cell proliferation and apoptosis assays; W.K. performed in situ PLA; S.L., Q.L., P.D. and S.T. performed all bioinformatics analysis; W.G.R. and K.-K.W. provided human lung cancer patient samples; N.L. provided eIF3 complex. P.Sl., P.Sa., R.E.G. and F.J.S. contributed to discussion. J.C., S.L. and R.I.G. analysed data and wrote the paper with input from other authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0538-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0538-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.I.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Cell culture and transfection. Human lung cancer cell lines (A549 and H1299), HEK293T, BJ, NIH-3T3, HeLa cells and MEFs were cultured with DMEM supplemented with 10% fetal bovine serum (FBS) and antibiotics. Cells were grown in a 5% CO_2 cell culture incubator at 37 °C. Cell lines were authenticated using morphology-, karyotyping- and PCR-based approaches by ATCC. The cell lines were tested for potential mycoplasma contamination and confirmed that they are mycoplasma negative. Transfection of plasmids was performed using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Down regulation of target genes by siRNA was performed using Lipofectamine RNAi Max (Invitrogen). The following siRNA sequences were used in this study: 5′-r(GAUAGAUGGCCUUGUGGUA)(UU)-3′ for eIF3h-1; 5′-r(GCGGAGCCU UCGCCAUGUA)(UU)-3′ for eIF3h-2; 5′-r(UGAGAAAGGAGGAGAGGAA) d(TT)-3′ for eIF4GI; 5′-r(UCAACCUCUUUACGGAUUU)d(TT)-3′ for eIF3b; and 5′-r(GCAUCAACCUGAAUGACAU)(UU)-3′ for CTIF.

Virus production and generation of stable knockdown and overexpression cells. Virus-mediated generation of stable knockdown and overexpression cells were performed as previously described[1]. In brief, shRNA containing pLKO.1 vector was co-transfected with pLP1, pLP2, and VSVG into 293T cells. For overexpression, pCDH vectors containing the wild-type *METTL3* (*METTL3* WT) and A155P cDNA were co-transfected with Delta 8.9 and VSVG plasmids into 293T cells. Viruses were collected at 48 h and 72 h after transfection and then used to infect cells with Polybrene (8 mg/ml, Sigma); 48 h after infection, puromycin was added to the culture medium to select the infected cells.

Plasmid construction. pFLAG-METTL3 WT, pFLAG-MS2-METTL3 and pFLAG-MS2-METTL3 Mut, have previously been described[1]. pFLAG-METTL3 A155P plasmids were generated by inducing point mutation in pFLAG-METTL3 WT using Q5 Site-Directed Mutagenesis Kit (NEB E0554). Plasmids pFLAG-MS2-METTL3 (1–100), pFLAG-MS2-METTL3 (1–150), pFLAG-MS2-METTL3 (1–200), pFLAG-MS2-METTL3 (1–350) and pFLAG-MS2-METTL3 (101–580) was constructed by substitution of PCR-amplified each METTL3 fragment into the NotI–*METTL3* WT–BglII site of pFLAG-MS2-METTL3. For stable METTL3 overexpression, wild-type *METTL3* and A155P sequence were PCR-amplified and cloned into the NheI and NotI sites of pCDH-CMV-MCS-EF1-Puro plasmid. For expression of recombinant METTL3 proteins, cDNA corresponding to full-length METTL3 and N-terminal amino acids 1–200 were cloned into the pETDuet-1 and pET His6 GST TEV LIC cloning vector individually. For expression of recombinant METTL3 proteins for in vitro translation, Flag–MS2, Flag–MS2–METTL3 (1–200), Flag–MS2–METTL3, Flag–MS2–METTL3(A155P) were cloned into pETDuet-1. For bacteria protein expression of human eIF3h, eIF3j and eIF3m that express N-terminal GST-fused proteins, BamHI/EcoRI fragment of pGEX2TK vector was ligated to the PCR amplified BamHI/EcoRI fragment that contained eIF3h, eIF3j or eIF3m. In addition, for the pGEX2TK-eIF3g or -eIF3i, BglII/EcoRI fragment of pGEX2TK vector was ligated to the PCR amplified BglII/EcoRI fragment that contained either eIF3g or eIF3i, respectively. For bacteria expression of eIF3h, deletion mutants were generated by ligation of BamHI/EcoRI fragment of pGEX2TK vector with BamHI/EcoRI fragment of either PCR-amplified eIF3h (1–222) or eIF3h (29–222). The pGL3c_TK luciferase reporter (FLuc) and pGL3c_TK luciferase reporter containing 2× MS2 binding sites near the stop codon (FLuc-MS2bs) have previously been described[1]. The 2× MS2 binding site sequence was PCR-amplified from FLuc-MS2bs and inserted into the NcoI site of pGL3c_TK luciferase reporter to make the FLuc-5′ UTR-MS2bs reporter that the MS2 binding sites are located in the 5′ UTR region of the luciferase gene. The 2× MS2 binding site sequence and GFP sequence (from CAG-GFP, Addgene Plasmid #16664) were cloned into the XbaI site of pGL3c_TK luciferase reporter (FLuc) to make the FLuc-MS2bs-GFP and FLuc-GFP-MS2bs reporters. For pFLAG-tethering effector plasmids, *METTL3* shRNA resistance plasmids were generated by introducing synonymous mutations into the shRNA targeting sequence using the Q5 Site-Directed Mutagenesis Kit (NEB E0554). All cloning primers are listed in Supplementary Table 1.

In vitro translation assay. H1299 cells were collected and resuspended in hypotonic buffer (10 mM Hepes (pH 7.4), 10 mM potassium acetate, 1.5 mM magnesium acetate, and 2.5 mM dithiothreitol (DTT)). Cells were then incubated on ice for 30 min and ruptured by passing ten times through a 25-gauge needle attached to a 3-ml syringe. The cell homogenate was centrifuged at 13,000*g* for 15 min at 4 °C. The supernatant was collected and used for in vitro translation. In vitro translation reactions were performed for 1 h at 30 °C using either H1299 cytoplasmic cell extracts or rabbit reticulocyte lysate (RRL) (Thermo FisherScientific, AM1200) in 20-μl reaction mixtures containing 100 ng of in vitro transcribed reporter mRNAs and 500 ng of either purified recombinant His–Flag–MS2, His–Flag–MS2–METTL3 or His–Flag–MS2–METTL3 (1–200) protein. The activity of in vitro-translated luciferase was measured using a Luciferase assay kit (Promega, E1960) according to the manufacturer's instructions. Reporter mRNAs either presence or absence of poly(A) tails were in vitro transcribed using PCR-amplified FLuc-MS2bs fragment with following primers; 5′-GACTAGTAATACGACTCACTATAGGG GCCACCATGGAAGACGCCAAAAACATAAAG-3′ (sense) and 5′-TCTAGA CCCCGGGAGCATGGGTGAT-3′ (antisense) for the FLuc-MS2bs Poly(A)$^-$ mRNA, and 5′-GACTAGTAATACGACTCACTATAGGGGCCACCATGGA AGACGCCAAAAACATAAAG-3′ (sense) and 5′-TTTTTTTTTTTTTTTTTTTT TTTTTTTTTTTCTAGACCCCGGGAGCATGGGTGAT-3′ (antisense) for the FLuc-MS2bs Poly(A)$^+$ mRNAs.

RNA isolation and RT–qPCR. The details of RNA isolation and RT–qPCR assays are as previously described[1]. In brief, RNA was extracted from cells, co-immunoprecipitation or sucrose gradient fractionation samples using Trizol (Invitrogen) according to the manufacturer's instructions. RT–qPCR analyses were performed using SYBR Green PCR Master Mix with the Step One Real-Time PCR System (AppliedBiosystems). All primers used in this study are listed in Supplementary Table 2. For the analysis of global or individual mRNA lifetime, METTL3-depleted or control HeLa cells (60-mm culture dishes) were treated with actinomycin D (5 μg/ml), then collected after 0 h, 2 h, 4 h and 6 h.

Luciferase assay and translation efficiency. Dual luciferase assays were performed according to the manufacturer's protocol (Promega). FLuc activity was normalized to the Renilla luciferase (RLuc) activity. Relative FLuc activity was normalized to the relative FLuc mRNAs. The normalized FLuc activity (translation efficiency) in the presence of Flag–MS2 was set to 1.

Polysome fractionation and RNA-seq. METTL3-depleted or control HeLa cells (four 150-mm culture dishes) were treated with 100 μg/ml cycloheximide (Sigma) for 10 min at 37 °C. Cells were then lysed and layered onto 10–50% sucrose gradient tube and centrifuged at 36,000 r.p.m. in a Beckman SW-41Ti rotor for 2.5 h at 4 °C. Gradients were fractionated and monitored at absorbance 254 nm (Brandel). Collected fractions were pulled into sub-polysome fraction and polysome fraction. Then, total RNA, sub-polysome and polysome samples were subjected to RNA-seq. Poly(A)-selected mRNAs were purified and used for library construction using TruSeq Stranded mRNA Sample Prep Kits (Illumina RS-122-2101) and sequenced with Illumina NextSeq 500. ERCC RNA Spike-In Control Mixes (Ambion) were added into each sample before constructing the library to normalize the reads.

Co-immunoprecipitation, mass spectrometry and western blot. Co-immunoprecipitation and western blot were performed as previously described[1]. In brief, Flag–METTL3-expressing HeLa or H1299 cells were collected and lysed using NET-2 buffer (50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 1 mM phenylmethanesulfonylfluoride (PMSF), 2 mM benzamidine, 1% NP-40) then the supernatant was subjected to immunoprecipitation using anti-Flag M2 Affinity Gel (Sigma-Aldrich). Where indicated, the affinity elute was subjected to SDS–PAGE and then either colloidal Coomassie blue staining or western blotting. Bands were excised and subjected to mass spectrometric sequencing as previously described[1]. The following antibodies were used for western blotting: METTL3 (Proteintech, 15073-1-AP; Abcam, ab195352), β-actin (Abcam, ab8227), eIF3h (Abcam, ab60942) CBP80 (gift from Y. K. Kim, Korea University), CTIF (gift from Y. K. Kim, Korea University), eIF4E (Cell Signaling Technology, #2067), eIF3b (Santa Cruz Biotechnology, sc-16377), eIF4GI (Cell Signaling Technology, #2498), Flag (Sigma, A8592), BRD4 (Abcam, ab128874), CD9 (Cell Signaling Technology, #13174), MGMT (Cell Signaling Technology, #2739), TIMP1 (Cell Signaling Technology, #8946) and FTO (Phosphosolution, 597-FTO).

Protein expression, purification and GST pull-down assay. Plasmids expressing the recombinant proteins were transformed into BL21 *Escherichia coli* and then the recombinant proteins were induced by IPTG at 20 °C overnight. The bacteria were pelleted and resuspended in protease inhibitor containing PBST buffer and then lysed by sonication. His-tag recombinant proteins were purified using Ni-NTA agarose (Qiagen 30210). The GST-tagged proteins were purified using the glutathione sepharose (BioVision 6655) according to the manufacturer's protocol. For GST pull down assay, equal amount of GST fusion proteins or GST control bound to glutathione sepharose were incubated with purified recombinant His-tagged full-length METTL3 or N-terminal (1–200) fragment for 1 h at 4 °C, after extensive washing, the proteins bound to the sepharose were resolved on SDS–polyacrylamide gels and detected by western blot analysis.

Far-western blotting. Far-western blotting was performed with biochemically purified human eIF3 protein complex that was resolved by SDS–PAGE and then transferred to Hybond ECL nitrocellulose membrane. The membrane was first incubated in blocking buffer (100 mM Tris (pH 7.5), 100 mM potassium acetate, 2 mM magnesium acetate, 0.1 mM EDTA, 10% glycerol, 1 mM PMSF, 1 mM benzamidine, and 0.05% Tween 20, 5% non-fat milk) at 4 °C overnight, then the membrane was incubated with blocking buffer containing 5 μg of purified recombinant full-length METTL3 or METTL3 (1–200) proteins at 4 °C for another 24 h. After that, the membrane was incubated with METTL3 antibody for western blotting analysis.

Electron microscopy. Flag–METTL3 expressing H1299 cells were collected and lysed using NET-2 buffer, then the supernatant was subjected to immunoprecipitation using anti-Flag M2 Affinity Gel (Sigma-Aldrich). Resin-bound mRNP complexes were eluted using 3×Flag peptides (Sigma, F3290). Where indicated, during the elution, anti-METTL3 antibody (Proteintech, 15073-1-AP) and gold nanoparticle

© 2018 Springer Nature Limited. All rights reserved.

(6 nm) conjugated anti-rabbit IgG were added with or without either anti-CBP80 antibody or anti-eIF4E antibody that was gold-nanoparticle (10 nm)-conjugated using GOLD conjugation kit (Abcam, ab201808), according to the manufacturer's instructions. The eluates were then fractionated using 10–50% sucrose gradients. Each fraction was applied to an electron-microscopy grid (EMS, G400-Cu) covered with a thin layer of carbon and after 1 min the excess suspension was blotted with a filter paper. The grid was washed twice with water and 0.7% uranyl formate, and then negatively stained for 20 s with 0.7% uranyl formate. The specimens were examined in Tecnai G2 Spirit BioTWIN Transmission Electron Microscope (FEI company) with AMT 2k CCD camera equipped. Direct magnification of 68,000× was used to detect images from Extended Data Fig. 9a, and magnification of 98,000× was used to detect images from Figs. 1b, 3h and Extended Data Fig. 4a, c. Polysome numbers were counted using 20 individual pictures for each sample with direct magnification of 30,000× in Fig. 3f. All the images shown in the figures are cropped sections. The average distance between immuno-gold particles in Extended Data Fig. 4d was measured from the images in Fig. 1b and Extended Data Fig. 4c using the electron microscope software (AMT Capture engine). In vitro translation reactions were performed for 1 h at 30 °C using RRL, and then the total reaction mixture was subjected to illustra MicroSpin S-400 HR Columns (Fig. 3f, Extended Data Fig. 9a) or sucrose gradient fractionation (Fig. 3g, h). The eluates were then applied to an electron-microscopy grid and analysed by electron microscopy.

Cap-association assay using m^{7}GTP-agarose. To analyse the interaction of METTL3 and the cap-binding protein complex, cells were lysed using NET2 buffer and total cell extracts were incubated with m^{7}GTP-agarose (Jena Bioscience, AC-155S) for 2 h at 4 °C. Then, the beads were washed five times and suspended in SDS sample buffer. The eluted samples were analysed by western blot. Where indicated, 75 μM of m^{7}G(5′)ppp(5′)G Cap Analogue (Ambion, AM8048) was added into the sample and incubated with m^{7}GTP-agarose.

Cell proliferation, apoptosis and invasion assays. Cell proliferation, apoptosis and invasion assays were performed as described[1]. In brief, for cell proliferation, 700 cells were seeded in a 96-well plate on day 0 with the pertinent treatment. JQ1 (500 nM) was used. Absorbances at 490 nm were measured using CellTiter 96 AQueous One Solution Cell Proliferation Assay kit (Promega) on day 2, day 4, and day 6 to measure the cellular proliferation. The numbers of apoptotic cells were quantified by flow cytometric assays using Annexin V-FITC Apoptosis Detection Kit (BioVision) five days after 500 nM of JQ1 treatment and cell seeding. Cell invasion was measured using BioCoat Matrigel Invasion Chamber (Corning) according to the manufacturer's instructions.

Soft agar colony formation assays. NIH-3T3, MEFs and MB352 cells at 30% confluence were infected with the lentivirus expressing indicated protein for 48 h in the presence of 8 μg/ml polybrene (Sigma). Two days after infection, puromycin was added to the medium at 2.5 μg/ml, and cells were selected for 1 week. Fifty thousand live NIH-3T3 cells, 100,000 live MEFs or 100,000 live MB352 (p53 null MEFs) cells were selected and mixed with 0.35% top-agar and were plated onto 0.6% base-agar in six-well plates. Twenty-five days (NIH-3T3) or thirty days (MEFS or MB352) after plating the cells into soft agar, colony numbers were counted. The colony numbers were counted by openCFU.

PLA. HeLa cells were incubated with primary antibodies (rabbit anti-METTL3 antibody and mouse anti-eIF3h antibody) in blocking solution at 4 °C for 2 h. Cells were then washed for five times for 5 min in PBS plus 0.1% Tween 20. Then, cells were incubated with secondary proximity probes (anti-Rabbit-PLUS and anti-Mouse-MINUS) (Sigma, DUO92101) for 90 min at 37 °C. Cells were washed five times for 5 min in 10 mM Tris-HCl (pH 7.5) plus 0.1% Tween 20 at 37 °C, then twice for 5 min in PBS plus 0.1% Tween 20. All subsequent steps were performed according to the manufacturer's instruction. Cells were observed with a Zeiss LSM 710 Multiphoton Laser Scanning Confocal.

Immunohistochemistry staining. The human lung cancer tumour array was purchased from Biomax (HLug-Ade150CS-01). The slide was baked for 60 min in an oven set to 60 °C and then loaded into the Bond III staining platform with appropriate labels. Antigen was retrieved by Bond Epitope Retrieval 2 for 20 min. Then the slide was incubated with METLL3 antibody (Abcam, ab195352) at 1:500 for 30 min at room temperature. Primary antibody was detected using Bond Polymer Refine Detection kit. Slides were developed in DAB, then dehydrated and coverslipped. Each sample was scored by the percentage of positively stained cells (percentage score: 1–5) and the staining intensity (intensity score: 1–5). Then the sample staining score was calculated by multiplying the percentage score and the intensity score.

In vivo tumour xenograft. All research involving animals complied with protocols approved by the Beth Israel Deaconess Medical Center Institutional Animal Care and Use Committee. Four-to-six-week-old female NU/J (Nude) immunodeficient mice (Jackson Laboratory #002019) were used for subcutaneous injections. Randomly divided five mice (A549 cells) or eight mice (NIH-3T3 cells) were used for each group. One hundred thousand A549 cells or 1,500,000 NIH-3T3 cells in serum-free medium and growth-factor-reduced Matrigel (Corning #354230) (1:1) were inoculated into the flank of nude mice. The xenograft tumour formation was monitored using callipers twice a week. The recipient mice were monitored and euthanized when the tumours reached 1-cm diameter. The tumour volume was calculated by use of a formula 1/2(length × width2). The investigator was blinded to group allocation.

m^{6}A meRIP–seq and data analysis. meRIP–seq was performed as previously described[1]. All the studies involving human patient samples complied with protocols approved by the Institutional Review Board. Informed consent was obtained from all participants. Portions of fresh tumour tissue approximately 0.5 × 0.5 × 0.5 cm were snap-frozen and preserved at −80 °C, pulverized in liquid nitrogen and stabilized in Trizol for total RNA isolation. Then, the mRNA purification from total RNA was performed using PolyATtract mRNA Isolation Systems (Promega). Two micrograms of the purified mRNA was fragmented and immunoprecipitated with anti-m^{6}A antibody (Synaptic Systems, 202003). The purified RNA fragments from m^{6}A MeRIP were used for library construction using the TruSeq Stranded mRNA Sample Prep Kits (Illumina RS-122-2101) and sequenced with Illumina NextSeq 500. Reads mapping, peak calling, metagene analysis and motif search were performed as previously described[1]. To identify the alternative splicing events, all the clean RNA-seq reads of control and METTL3 knockdown samples were first trimmed to the same length (72 bp), which were then aligned against the human hg19 (GRCh37) reference genome using Tophat2[24]. rMATS v.3.2.5[25] was used to detect the splicing events and significant splicing differences between METTL3 knockdown and control samples. To analyse the global profiling of mRNA lifetime, the clean reads were aligned to human reference genome (hg19) using Tophat2[24] after trimming the adapters and filtering low-quality sequences from the raw data. The reads mapped to each gene were counted using HTSeq[26] on the basis of the GENCODE gene model (v19)[27]. The raw counts were then normalized as reads per kilobase per million mapped reads. ERCC RNA Spike-In Control Mixes (Ambion) were added into each sample before constructing the library to normalize the reads. mRNA lifetime was calculated according to the method in the previous study[8]. To analyse METTL3 or eIF3h expression level among tumours from the Cancer Genome Analysis (TCGA), RNA-seq data for 33 TCGA tumour types were downloaded from Genomic Data Commons Data Portal (GDC) of TCGA (https://cancergenome.nih.gov/) using R package TCGAbiolinks[28] The expression matrix was then constructed by merging the transcripts per million values of all downloaded RNA-seq samples. The tumour types without corresponding normal tissue samples were excluded and the retained 24 tumour types with normal tissues were used to draw the box plot of gene expression. *P* values are calculated by Wilcoxon rank-sum test, with asterisks indicating statistical significance.

Statistics and reproducibility. Data are presented as the mean ± s.e.m. or mean ± s.d. Statistical significance was determined by a Student's two-tailed *t*-test for RT–qPCRs and Luciferase assays. Where applicable, Shapiro–Wilk test and Q–Q plotting (quantile–quantile plot) were performed before Student's two-tailed *t*-test to assess whether the data plausibly came from a normal distribution. Where indicated, Shapiro–Wilk signed-rank test was used. $P < 0.05$ was considered statistically significant. Sample size for Extended Data Fig. 10a, b; $n = 408$, $n = 1{,}095$, $n = 304$, $n = 36$, $n = 285$, $n = 184$, $n = 155$, $n = 520$, $n = 66$, $n = 533$, $n = 290$, $n = 371$, $n = 515$, $n = 501$, $n = 178$, $n = 179$, $n = 497$, $n = 94$, $n = 259$, $n = 103$, $n = 415$, $n = 505$, $n = 120$ and $n = 176$, from the order of left to right for primary solid tumour; $n = 19$, $n = 133$, $n = 3$, $n = 9$, $n = 41$, $n = 11$, $n = 5$, $n = 44$, $n = 25$, $n = 72$, $n = 32$, $n = 50$, $n = 59$, $n = 51$, $n = 4$, $n = 3$, $n = 52$, $n = 10$, $n = 2$, $n = 1$, $n = 35$, $n = 59$, $n = 2$ and $n = 24$, from the order of left to right for normal tissue. Sample size for Extended Data Fig. 10c; $n = 285$, $n = 184$, $n = 155$, $n = 371$, $n = 515$, $n = 501$, $n = 497$ and $n = 94$, from the order of left to right.

Code availability. Script and code used for data analysis can be found at https://github.com/rnabioinfor/rnamethy.

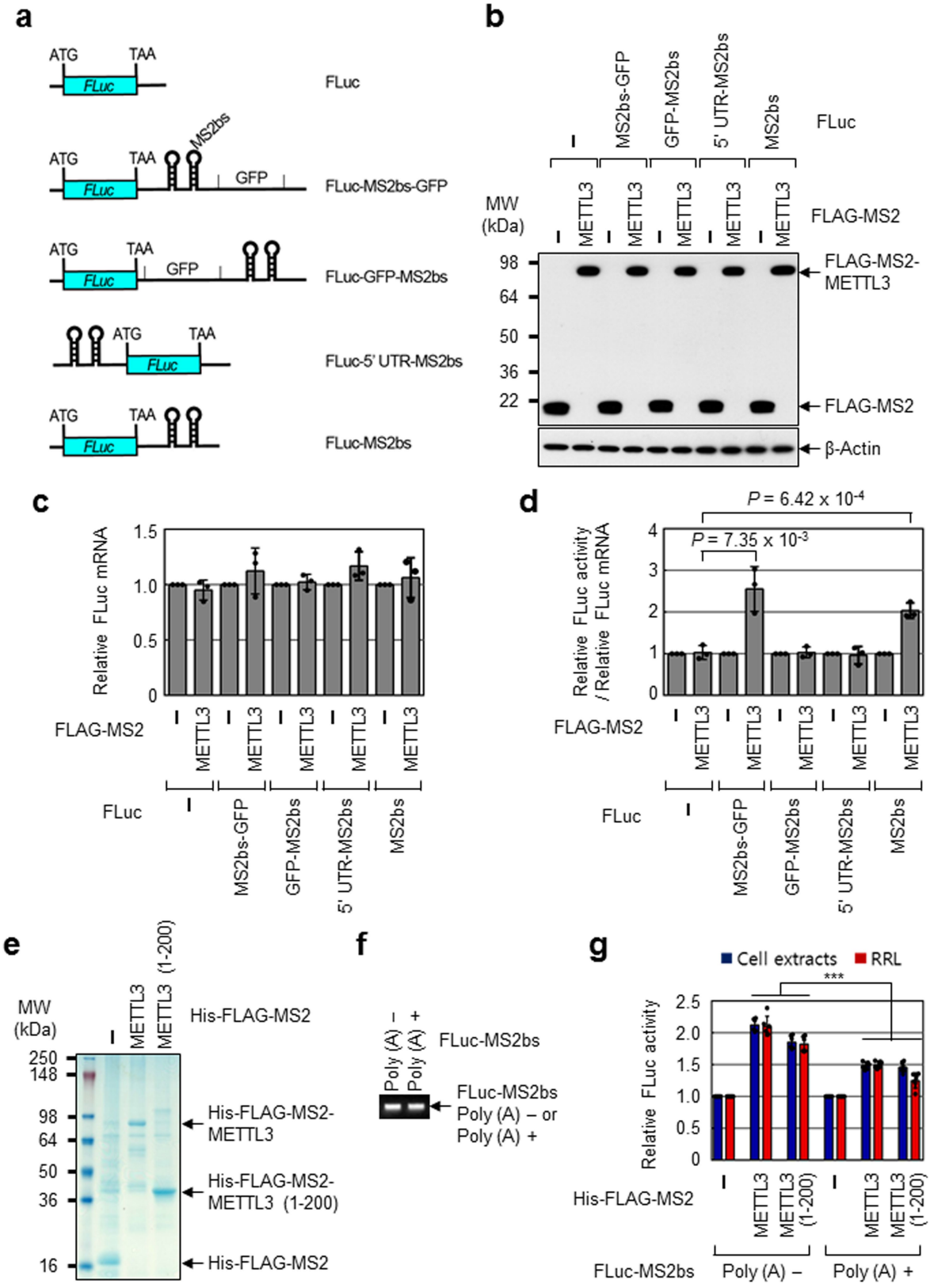
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The m^{6}A meRIP–seq and RNA-seq data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE117299. All other data are available from the authors on request.

24. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
25. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
26. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
27. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
28. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).

© 2018 Springer Nature Limited. All rights reserved.

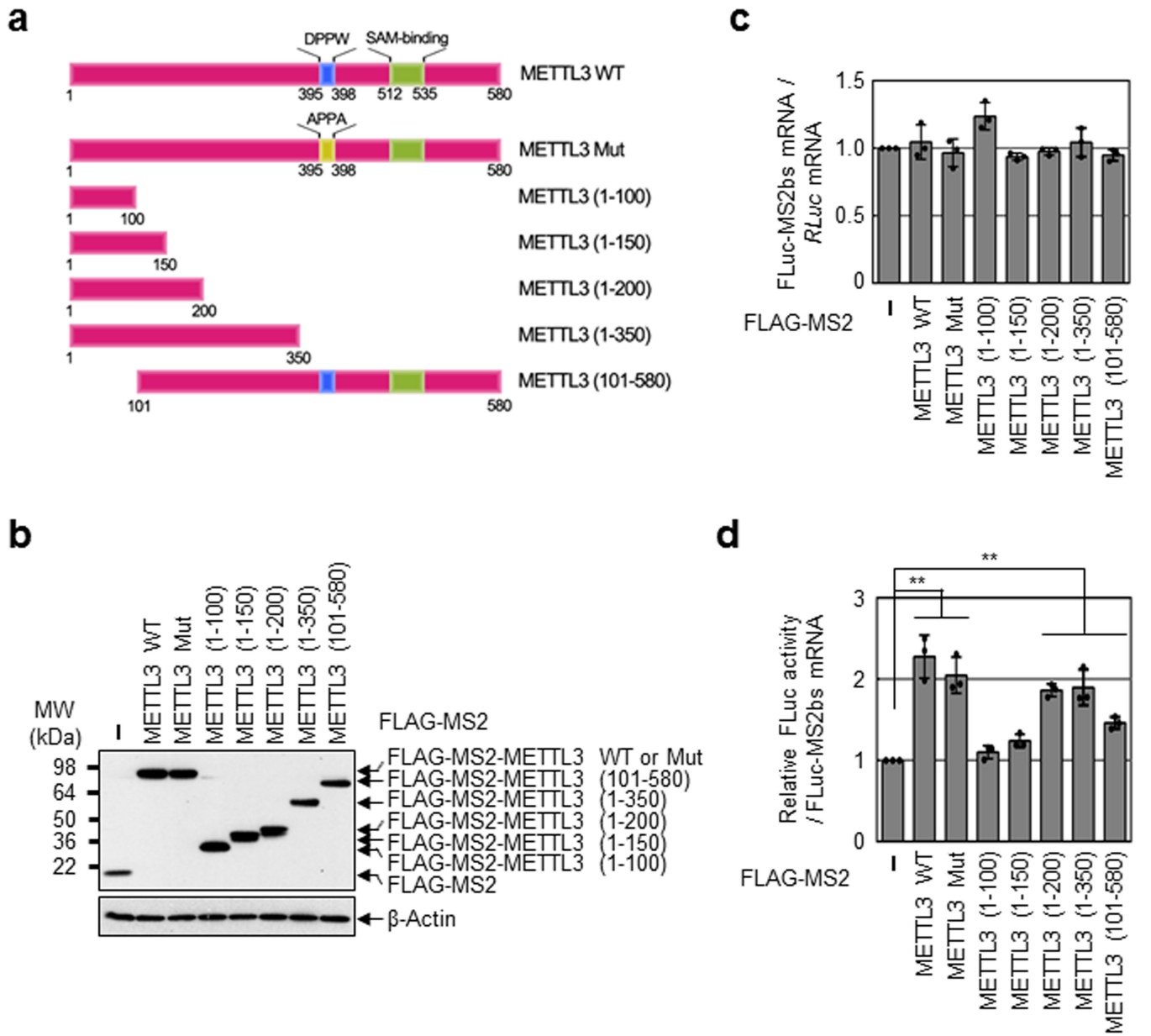


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

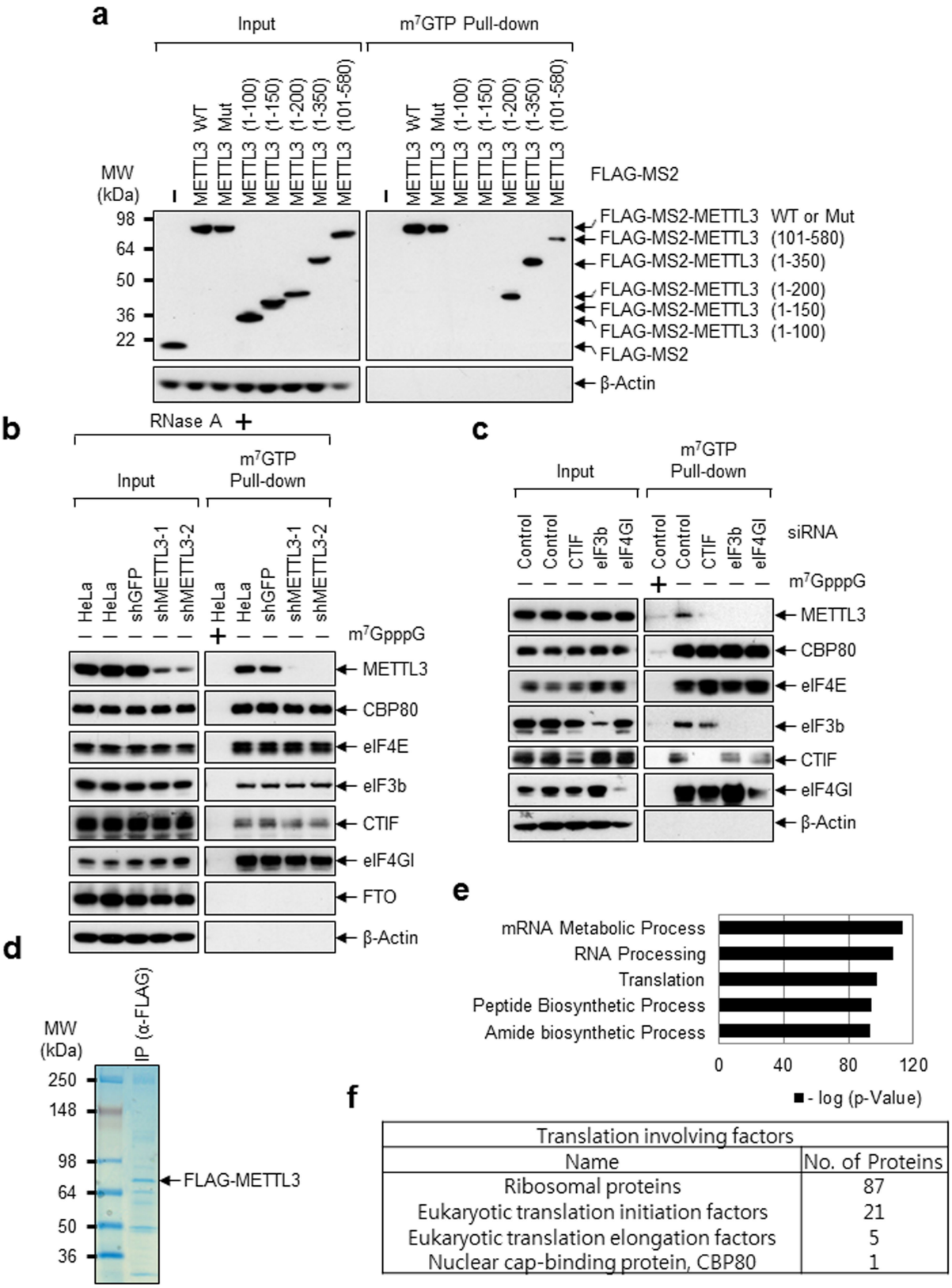
Extended Data Fig. 1 | METTL3 binding close to the stop codon enhances translation. a, Schematic of reporter plasmids containing Firefly luciferase cDNA and different positions of MS2 binding sites. **b**, Western blotting with indicated antibodies. Blot is representative of two independently performed experiments with similar results. **c**, RT–qPCR analysis of reporter mRNAs. Each tested reporter mRNAs were normalized to RLuc mRNAs. The FLuc:RLuc ratio for each construct with Flag–MS2 expression was set to 1. Data are mean ± s.d. from three biologically independent samples. **d**, Tethering assay to measure translation efficiency as described in Fig. 1h. Data are mean ± s.d. from three biologically independent samples. *P* values from a two-sided *t*-test. **e**, Colloidal Coomassie blue staining of recombinant protein His–Flag–MS2, His–Flag–MS2–METTL3, or His–Flag–MS2–METTL3 (1–200). **f**, Ethidium bromide-stained agarose gel electrophoresis of the indicated in vitro transcribed reporter mRNAs; FLuc-MS2bs without poly (A) tail (Poly (A) -) or FLuc-MS2bs with 30-nt poly (A) tail (Poly (A) +). **f**, **g**, Images are representative of two independently performed experiments with similar results. **g**, In vitro translation of reporter mRNAs using either H1299 cell extracts or rabbit reticulocyte lysate (RRL). The levels of in vitro-translated FLuc protein were analysed using luciferase assays. Value of FLuc activity in the presence of His–Flag–MS2 recombinant protein was set to 1. Data are mean ± s.d. from six independent experiments. *P* value is from a two-sided *t*-test; $***P < 0.001$, multiple comparison.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | N-terminal region of METTL3 promotes translation. a, Schematic of METTL3-deletion mutants or mutation in METTL3 catalytic domain. **b**, Western blotting with indicated antibodies. The blot is representative of two independently performed experiments with similar results. **c**, RT–qPCR analysis of reporter mRNAs. FLuc–MS2bs mRNA levels were normalized to RLuc mRNAs. The FLuc:RLuc ratio obtained in Flag–MS2 (control) was set to 1. Data are mean ± s.d. from three biologically independent samples. **d**, Tethering assay to measure translation efficiency of reporter mRNAs as described in Fig. 1h. Data are mean ± s.d. from three biologically independent samples. *P* value is from a two-sided *t*-test; $^{**}P < 0.01$, multiple comparison.

© 2018 Springer Nature Limited. All rights reserved.



f

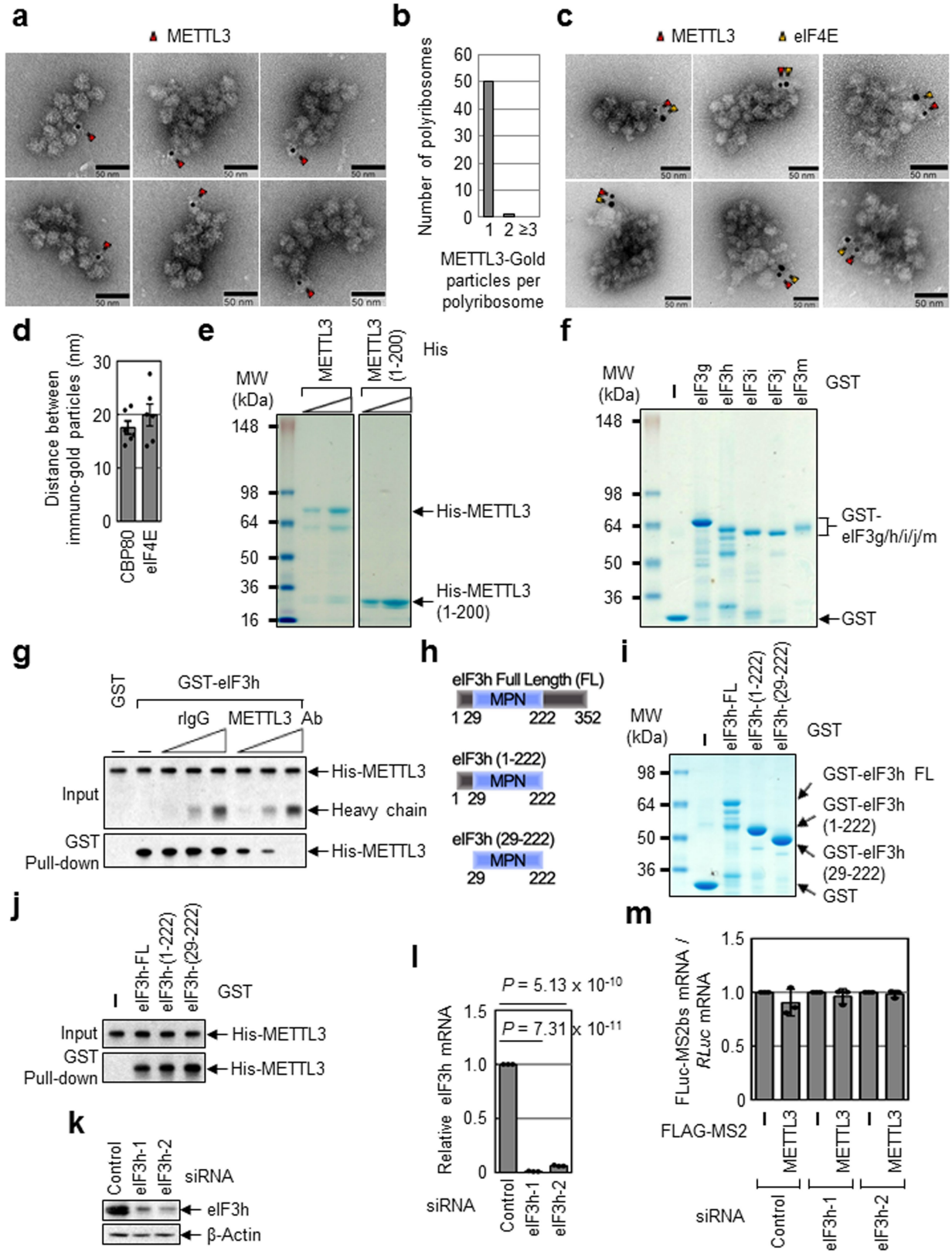
Translation involving factors	
Name	No. of Proteins
Ribosomal proteins	87
Eukaryotic translation initiation factors	21
Eukaryotic translation elongation factors	5
Nuclear cap-binding protein, CBP80	1

Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 3 | METTL3 associates with translation initiation factors. a, Deletion mutants of METTL3 were expressed in HeLa cell. The total-cell extracts (Input) and the cap-associated protein samples were analysed by western blotting using the indicated antibodies. **b**, Cap-association assay with METTL3 depletion. The total-cell extracts (Input) and the cap-bound protein samples were analysed by western blotting using the indicated antibodies. m^7GpppG cap analogue was used for antagonizing cap-associating proteins binding to m^7GTP-agarose. **c**, Same as **b**, except HeLa cells were transfected with *CTIF*, *EIF3B* or *EIF4G1* siRNA. In **a**, **b**, blots are representative of two independently performed experiments with similar results. **d–f**, Mass spectrometry of Flag–METTL3 interacting proteins. **d**, Proteins that were co-immunopurified with Flag–METTL3 subjected to 4–12% Tris–glycine SDS–PAGE. Colloidal Coomassie blue staining was performed. **e**, Gene Ontology analysis of the identified proteins from mass spectrometry. One experiment was performed. Hypergeometric distribution (one-tail) with Bonferroni adjustment was used to determine enrichment statistical significance. **f**, Table showing the translation involving factors identified from mass spectrometry.

© 2018 Springer Nature Limited. All rights reserved.

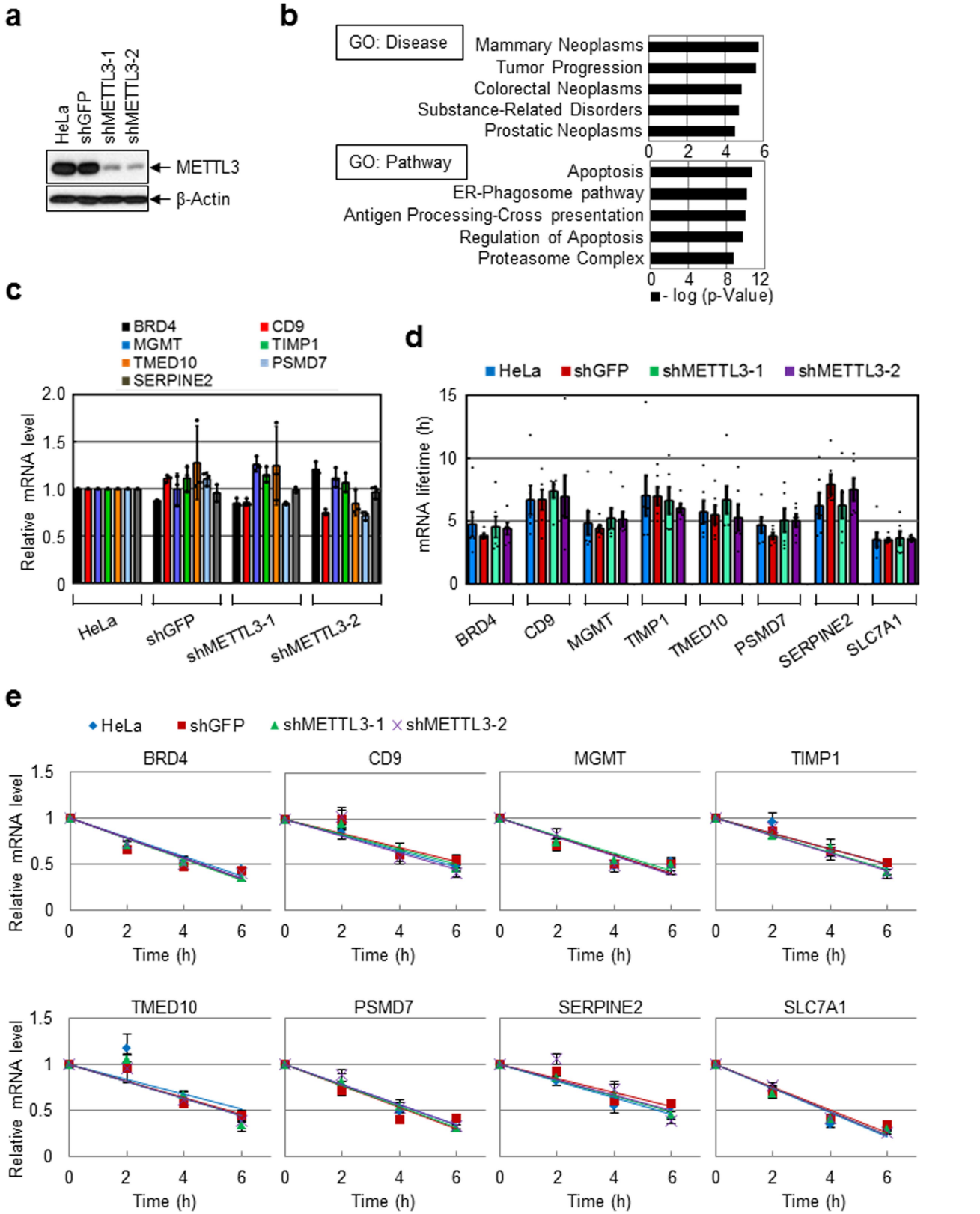


Extended Data Fig. 4 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

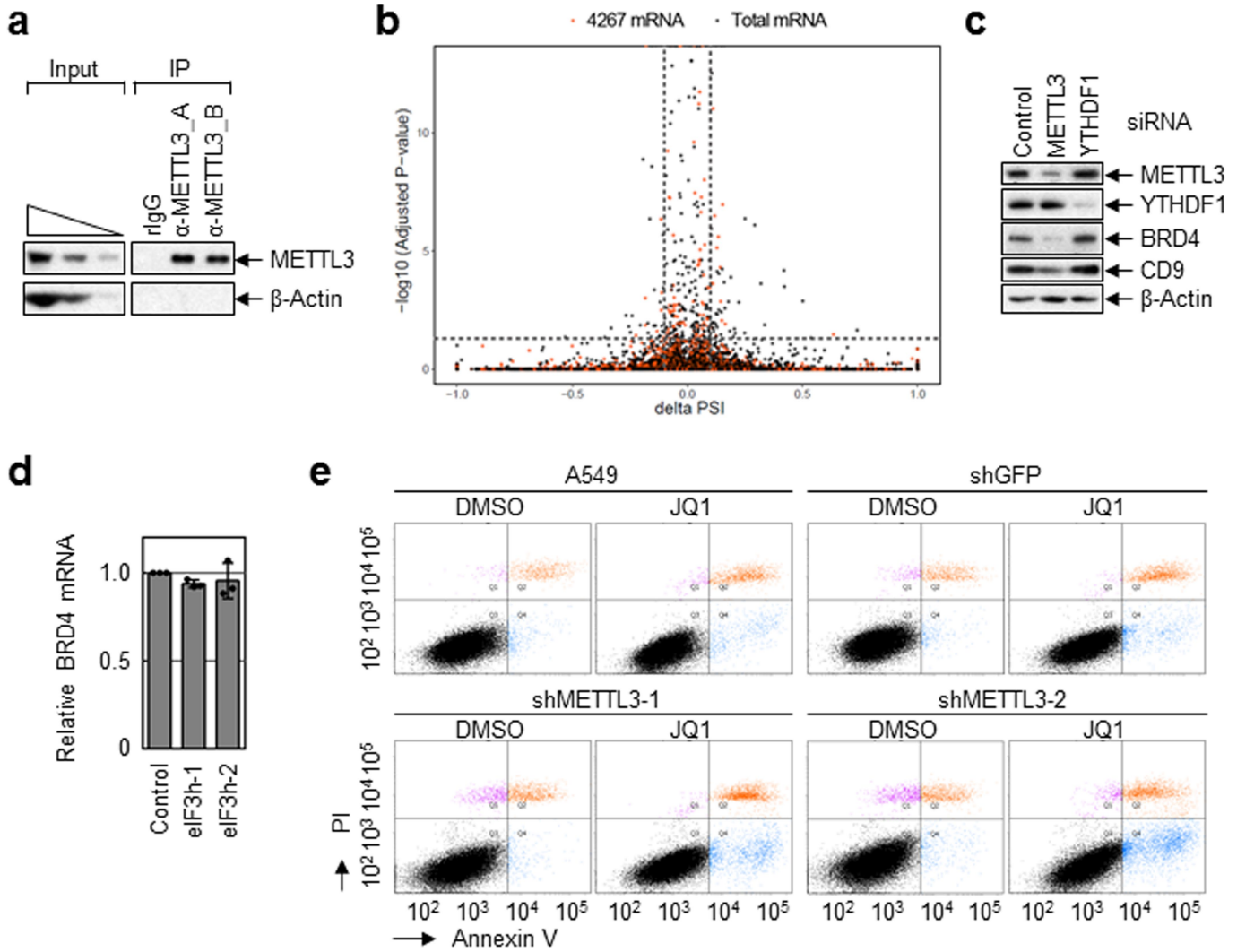
Extended Data Fig. 4 | N-terminal region of METTL3 directly interacts with MPN domain of eIF3h. a, Electron-microscopy images of polyribosome with gold-particle labelling of METTL3. Red arrows indicate METTL3 with immuno-gold particle (6 nm). Images are representative of three independent experiments with similar results. **b**, Counting of METTL3 with gold-particle labelling in each polyribosome. **c**, Electron-microscopy images of polyribosome with METTL3 and eIF4E. Red arrows indicate METTL3 with immuno-gold particle (6 nm) and yellow arrows indicate eIF4E with immuno-gold particle (10 nm). Images are representative of four independently performed experiments with similar results. **d**, Mean distance between immuno-gold particles was measured. Data are mean ± s.d. of six biologically independent samples from at least three independent experiments. **e**, Colloidal Coomassie blue staining of recombinant protein His–METTL3 or His–METTL3 1–200 amino acid fragments. **f**, Colloidal Coomassie blue staining of recombinant GST-tagged protein eIF3g, eIF3h, eIF3i, eIF3j or eIF3m. **g**, GST–eIF3h was co-purified with His–METTL3 in the presence of either rabbit IgG (rIgG) or anti-METTL3 antibody. Levels of co-purified His–METTL3 were analysed by western blotting. **e**–**g**, Images are representative of two independently performed experiments with similar results. **h**, Schematic diagram of human eIF3h deletion mutants. **i**, Colloidal Coomassie blue staining of recombinant GST–eIF3h, GST–eIF3h (1–222) or GST–eIF3h (29–222). **j**, GST pull-down of the indicated eIF3h deletion mutants. Co-purified His–METTL3 was analysed by western blotting. **i**, **j**, One experiment was performed. **k**, Western blotting demonstrates efficient knockdown of eIF3h protein. Blots are representative of three independently performed experiments with similar results. **l**, RT–qPCR analysis demonstrates efficient down regulation of eIF3h mRNA. Data are mean ± s.d. from three biologically independent samples. *P* values are from a two-sided *t*-test. **m**, RT–qPCR analysis of reporter mRNAs. FLuc–MS2bs reporter mRNAs were normalized to RLuc mRNAs. The FLuc:RLuc ratio obtained in Flag–MS2 was set to 1. Data are mean ± s.d. from three biologically independent samples.

© 2018 Springer Nature Limited. All rights reserved.



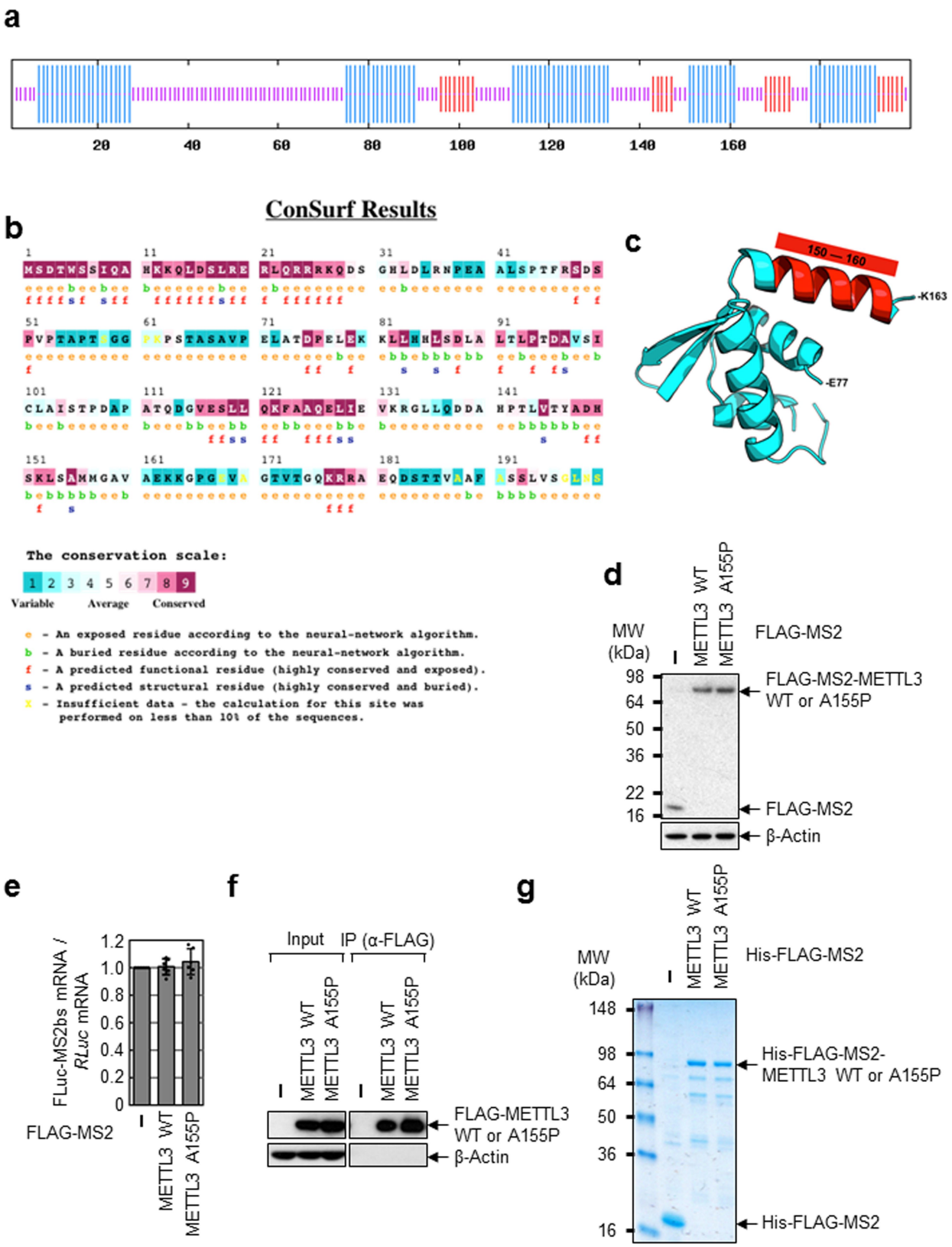
Extended Data Fig. 5 | METTL3 has no significant effect on mRNA stability. a, Western blotting with indicated antibodies. Blot is representative of three independently performed experiments with similar results. **b**, Gene Ontology analysis of the overlapping mRNAs ($n = 809$) in Fig. 2d. Hypergeometric distribution (one-tail) with Bonferroni adjustment was used to determine statistical significance of enrichment. **c**, RT–qPCR analysis using indicated primers. Data are mean ± s.d. from three technical replicates. **d**, **e**, Half-life of endogenous mRNAs was analysed by RT–qPCR using indicated primers. Data are mean ± s.e.m. from six independent experiments. **d**, *P* values are from a two-sided *t*-test; multiple comparison for the *P* values showed that there were no significant differences between the samples for all the tested mRNAs ($P > 0.05$).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Widespread role of METTL3 in oncogene translation. a, Immunoprecipitation of endogenous METTL3 and western blotting using the indicated antibodies. The blot is representative of two independently performed experiments with similar results. **b**, Density plot reflects the distribution of changes in percentage spliced in ΔPSI values and corresponding *P* values for alternative splicing events detected by rMATs v.3.2.5 (rMATs is developed on the basis of a hierarchical framework and likelihood-ratio test was used to detect differential splicing). Splicing events at a FDR < 5% and ΔPSI > 0.1 are considered as significant. Black dots indicate total mRNAs. Red dots (4,276 mRNAs) indicate mRNAs in METTL3-depleted cells that are translated more than twofold less. **c**, Western blot using indicated antibodies in control-, METTL3- or YTHDF1-knockdown cells. Blots are representative of two independently performed experiments with similar results. **d**, RT–qPCR analysis of endogenous *BRD4* mRNAs. Data are mean ± s.e.m. from three biologically independent samples. **e**, Annexin V/PI staining of METTL3 knockdown and control A549 cells upon JQ1 treatment (analysed by FACS). Data are from three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.

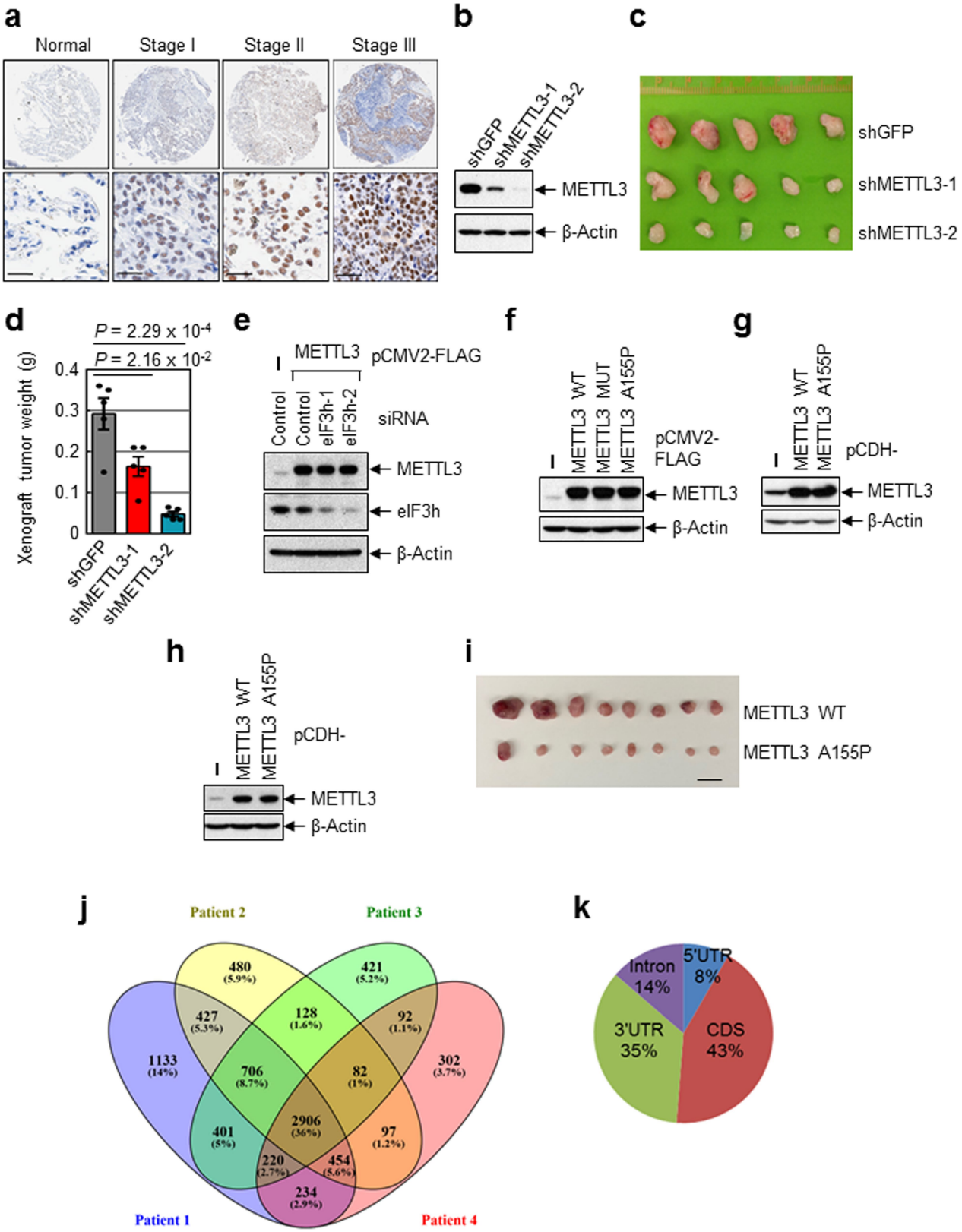


Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

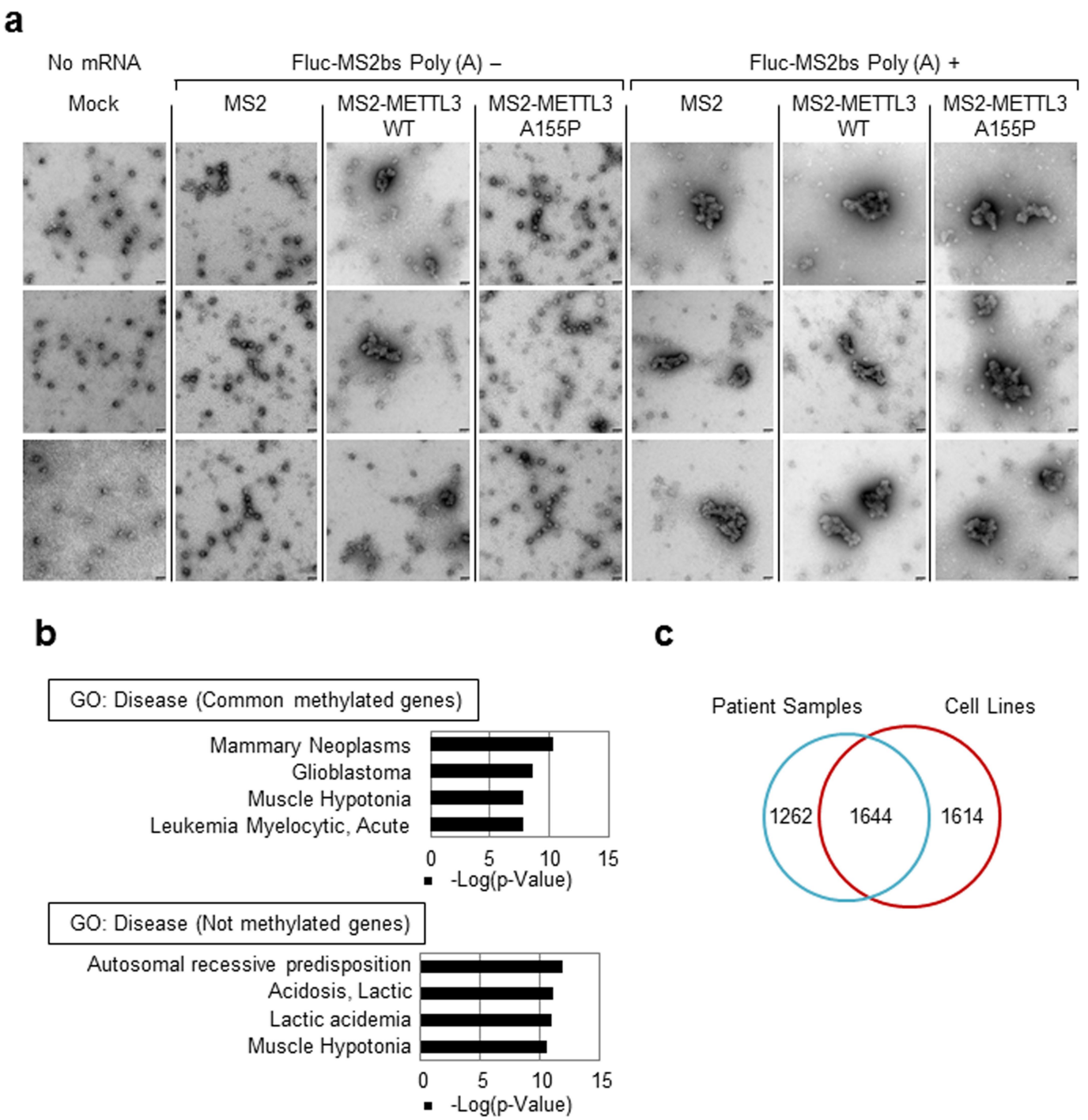
Extended Data Fig. 7 | Identification of a conserved alanine residue in the N-terminal region of METTL3 required for its interaction with eIF3h. a, Secondary structure prediction of the N-terminal (1–200) region of METTL3 protein showing putative alpha helices (blue lines). **b**, Evolutionary conservation of the N-terminal (1–200) region METTL3 protein. **c**, Computational modelling of the 3D structure of the N-terminal (77–163) region METTL3 protein, on the basis of the coordinates of RCSB Protein Data Bank entry 3HHH. **d**, Western blotting using the indicated antibodies. The blot is representative of two independently performed experiments with similar results. **e**, RT–qPCR analysis of reporter mRNAs. FLuc–MS2bs mRNA levels were normalized to RLuc mRNAs. The FLuc:RLuc ratio obtained in Flag–MS2 (control) was set to 1. Data are mean ± s.d. from six independent experiments. **f**, Immunoprecipitation of Flag–METTL3 (wild-type or METTL3(A155P)) and western blotting analysis using the indicated antibodies. Blots are representative of two independently performed experiments with similar results. **g**, Staining of recombinant protein wild-type His–Flag–MS2–METTL3 or His–Flag–MS2–METTL3(A155P). Gel is representative of two independently performed experiments with similar results.

© 2018 Springer Nature Limited. All rights reserved.



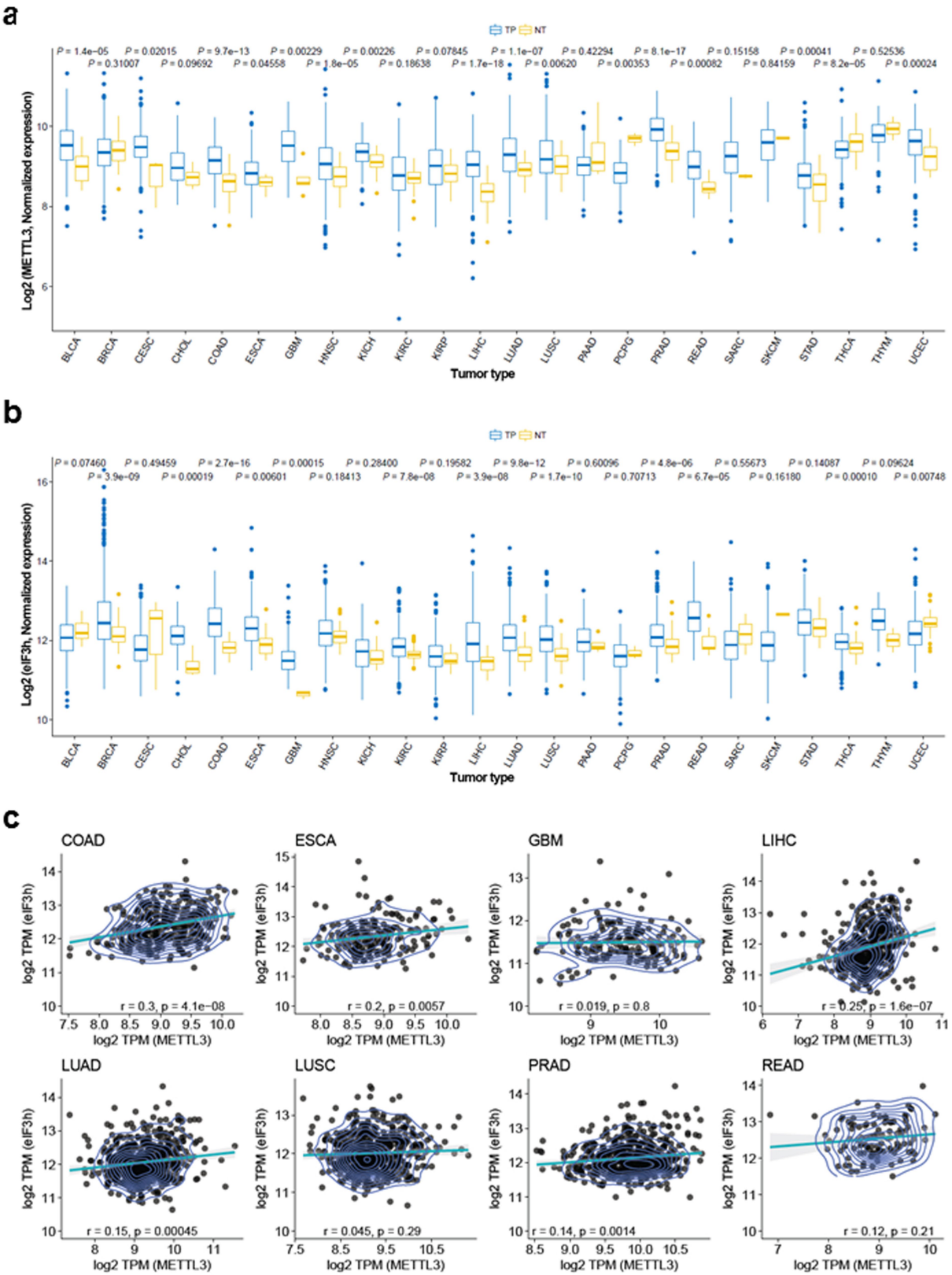
Extended Data Fig. 8 | METTL3 expression correlates with lung tumour stage and promotes tumorigenicity. a, Staining image of control and different stages lung cancer samples ($n = 75$). Bottom panels show the enlarged sections of the top panels. Scale bar, 30 μM. **b**, Western-blotting analysis using indicated antibodies. One experiment was performed. **c**, **d**, Tumour images (**c**) and plot of tumour weight (**d**) at the end point of the xenograft experiment. Data are mean ± s.e.m. from five independent mice. *P* values from a two-sided *t*-test. **e**–**h**, Western blotting analysis using the indicated antibodies. For **e**, two independently were performed experiments with similar results. For **f**–**h**, one experiment was performed. **i**, Tumour images at the endpoint in the xenograft experiment. Scale bar, 20 mm. **j**, Overlapping of genes containing m^6A identified in four samples from patients with lung cancer. **k**, Distribution of m^6A sites.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Polysome conformation is affected by METTL3 and m^{6}A modification in primary human lung tumours. a, Electron-microscopy images of polyribosomes. Images were taken from the samples in Fig. 3e. Scale bar, 50 nm. Six independently performed experiments show similar results. **b**, Gene Ontology analysis. Common methylated genes refers to the methylated genes in all four patient samples. 'Not methylated genes' indicates genes that are not methylated in any of the four patient samples. Hypergeometric distribution (one-tail) with Bonferroni adjustment was used to determine enrichment statistical significance. **c**, Venn diagram showing m^{6}A peak overlap between patient tumour samples and cells (H1299 and A549).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Expression of METTL3 and eIF3h is positively correlated in many tumour types. a, *METTL3* gene expression among TCGA tumours. **b**, eIF3h gene expression among TCGA tumours. **a**, **b**, Box plots display the full range of variation on the basis of the five number summaries (minimum, first quartile, median, third quartile, and maximum). NT, solid tissue normal; TP, primary solid tumour. Two-sided Wilcoxon signed-rank test was used for statistical significance. **c**, Plot illustrating the Pearson's correlations of expression level between METTL3 and eIF3h in eight TCGA tumours, in which both METTL3 and eIF3h are significantly changed when compared with normal tissues.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0526-z

Structure of the membrane–assembled retromer coat determined by cryo–electron tomography

Oleksiy Kovtun[1,2,8], Natalya Leneva[3,4,8], Yury S. Bykov[1,2], Nicholas Ariotti[3,6], Rohan D. Teasdale[3,7], Miroslava Schaffer[5], Benjamin D. Engel[5], David. J. Owen[4]*, John A. G. Briggs[1,2]* & Brett M. Collins[3]*

Eukaryotic cells traffic proteins and lipids between different compartments using protein-coated vesicles and tubules. The retromer complex is required to generate cargo-selective tubulovesicular carriers from endosomal membranes[1–3]. Conserved in eukaryotes, retromer controls the cellular localization and homeostasis of hundreds of transmembrane proteins, and its disruption is associated with major neurodegenerative disorders[4–7]. How retromer is assembled and how it is recruited to form coated tubules is not known. Here we describe the structure of the retromer complex (Vps26–Vps29–Vps35) assembled on membrane tubules with the bin/amphiphysin/rvs-domain-containing sorting nexin protein Vps5, using cryo-electron tomography and subtomogram averaging. This reveals a membrane-associated Vps5 array, from which arches of retromer extend away from the membrane surface. Vps35 forms the 'legs' of these arches, and Vps29 resides at the apex where it is free to interact with regulatory factors. The bases of the arches connect to each other and to Vps5 through Vps26, and the presence of the same arches on coated tubules within cells confirms their functional importance. Vps5 binds to Vps26 at a position analogous to the previously described cargo- and Snx3-binding site, which suggests the existence of distinct retromer-sorting nexin assemblies. The structure provides insight into the architecture of the coat and its mechanism of assembly, and suggests that retromer promotes tubule formation by directing the distribution of sorting nexin proteins on the membrane surface while providing a scaffold for regulatory-protein interactions.

Retromer forms complexes involved in different cellular trafficking routes by associating with sorting nexin (SNX) proteins that contain phox homology (PX) and bin/amphiphysin/rvs (BAR) domains (Vps5 and Vps17 in yeast), or with SNX proteins that lack BAR domains such as Snx3 or Snx27[8–12]. BAR domains form antiparallel homo- and heterodimers that are able to sense and/or modulate membrane curvature[13,14]. Retromer has previously been described as the cargo-sorting complex[15,16], although recent studies suggest that the retromer-associated SNX proteins are essential mediators of cargo binding[10,17].

Vps26–Vps29–Vps35 retromer heterotrimers were co-expressed in *Chaetomium thermophilum* (Extended Data Fig. 1a) and co-assembled with recombinant homodimeric Vps5 in solution (Fig. 1a, Extended Data Fig. 1b–d). Coated membrane tubules were produced in vitro by incubating retromer–Vps5 complexes with Folch lipid liposomes, yielding tubules with a membrane-to-membrane diameter of 31 ± 6 nm ($n = 209$). The sides and tips of tubules, as well as residual donor liposomes, were decorated with a protein coat of around 15-nm thick. The recruitment of retromer to the membrane was dependent on Vps5, whereas stable membrane tubulation by Vps5 was enhanced by retromer (Extended Data Fig. 1e, f). We imaged the resulting retromer–Vps5-coated tubules by cryo-electron tomography (cryo-ET) (Fig. 1b, Supplementary Videos 1, 2, 3), and applied reference-free subtomogram averaging to determine the structure and arrangement of the coat. Initial alignment revealed the presence of arch-like units on the membrane surface, which were subjected to local alignment procedures to generate overlapping density maps at subnanometer resolutions that together covered the complete, assembled structure (Fig. 1c, Extended Data Figs. 2, 3, Extended Data Table 1, Supplementary Video 4). Consistent with the resolution, α-helices were resolved in the cryo-ET maps. We determined the crystal structure of *C. thermophilum* Vps29 at 1.8 Å resolution (Extended Data Fig. 4a, Extended Data Table 2), and generated models of all other subunits of the protein coat from crystal structures of human homologues. These structures could be unambiguously modelled within the cryo-ET density maps to generate a pseudo-atomic structure of the assembled coat (Fig. 1c, d, Extended Data Fig. 5a–c, Supplementary Video 4; see Methods). Vps5 forms a pseudo-helical inner array on the membrane surface. Dimers of Vps26 interact directly with Vps5 and provide docking sites for Vps35 subunits that project away from the membrane. As observed in previous studies[17–19], the Vps35 α-solenoids form highly extended structures. Two copies of Vps35 meet at their C termini to form an arch, with Vps29 residing at the apex. By analogy, Vps26 forms the 'foot' between each Vps35 leg and the Vps5 layer. Retromer itself makes no direct contact with the membrane surface. Flexibility throughout the retromer–Vps5 coat enables it to adapt to differing membrane curvatures (Extended Data Fig. 5d, e, Supplementary Video 5).

Vps5 forms curved homodimers similar to those of Snx9 and Snx33[14] (Fig. 2a, b, Extended Data Fig. 6a). Vps5 dimers bind with their concave surface around the circumference of the tubule, similar to the binding of F-BAR and N-BAR domains[13,20], and contact the membrane through conserved, positively charged surfaces at the ends of the BAR domains as well as the PX domains (Fig. 2a, b). Vps5 dimers interact with each other to form rows through tip-to-tip contacts between the N-terminal end of the α1 helix of one BAR domain and the α2 and α3 helices of the next BAR domain (Fig. 2c). Lateral contacts between rows occur through the PX domains, via the loop between the first and second β-strands and the beginning of the second α-helix (Fig. 2c). Vps17 is likely to be structurally homologous to Vps5, in which case a heterodimeric array of Vps5 and Vps17 would form through equivalent contacts (Extended Data Fig. 6b, c).

The Vps26 subunit forms the sole contact between retromer and the membrane-proximal Vps5 array. Vps26 forms a homodimer via β-sheet complementation between the β7 strands of the N-terminal domains (Extended Data Fig. 4b). Each Vps26 dimer contacts four Vps5 dimers and bridges two adjoining Vps5 rows (Fig. 2a). The N-terminal subdomain of Vps26 has two electropositive loops, L5 (^{54}KDGKR^{58}) and L9 (^{134}RRMAD^{138}), which had previously been predicted to have a role in membrane association[17]. However, in the retromer–Vps5 complex, this site instead forms a complementary electrostatic cleft, engaging two Vps5 dimers near their tip-to-tip contacts (Fig. 2a).

[1]MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Cambridge, UK. [2]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [3]Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Queensland, Australia. [4]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [5]Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, Martinsried, Germany. [6]Present address: Electron Microscope Unit, The University of New South Wales, Kensington, New South Wales, Australia. [7]Present address: School of Biomedical Sciences, The University of Queensland, St. Lucia, Queensland, Australia. [8]These authors contributed equally: Oleksiy Kovtun, Natalya Leneva. *e-mail: djo30@cam.ac.uk; jbriggs@mrc-lmb.cam.ac.uk; b.collins@imb.uq.edu.au

© 2018 Springer Nature Limited. All rights reserved.

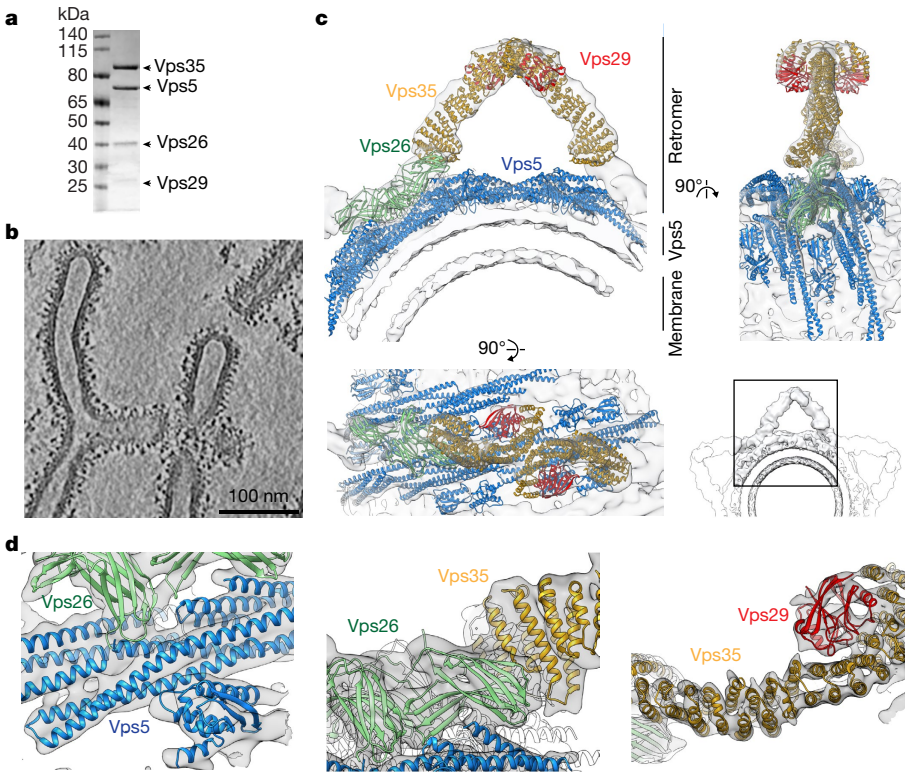


Fig. 1 | Cryo-ET structure of the membrane-associated retromer–Vps5 complex. **a**, Coomassie-stained SDS–PAGE gel of the purified retromer–Vps5 complex. **b**, Section through a cryo-electron tomogram of retromer–Vps5-coated membrane tubules. Panels **a** and **b** are representative of at least three independent experiments. **c**, Ribbon model of the retromer–Vps5 complex superimposed on overlapped, low-resolution electron density maps from an intermediate subtomogram alignment (Extended Data Fig. 2e). On the bottom right, three copies of the same density (one is boxed) are placed at positions related by the two-fold dimeric interface formed by Vps26, illustrating how the coat can propagate around the tubule. **d**, Close-up views of the retromer model fitted into the final high-resolution density maps.

A further electrostatic contact with the C-terminal domain of Vps26 induces movement of the C-terminal α-helix of one Vps5 monomer away from the rest of the coiled-coil bundle towards Vps26 (Fig. 2d), where it occludes the putative cargo-binding pocket identified in the human VPS26–VPS35–SNX3–DMT1-II peptide complex (Extended Data Fig. 4c, d). This is consistent with findings that Snx3-containing retromer coats may function in distinct cargo-sorting events from coats involving BAR-domain-containing SNX proteins[9,12]. Therefore, cargo

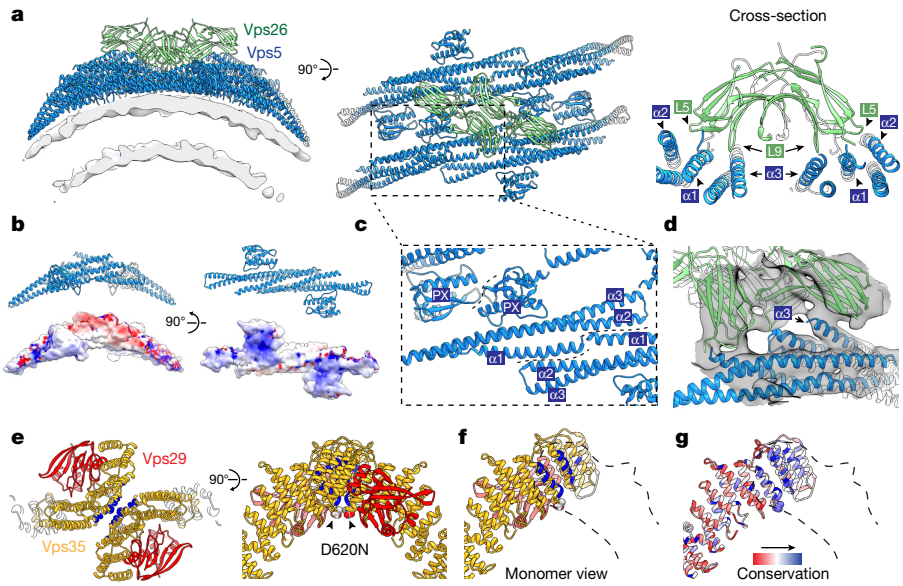


Fig. 2 | Structures of interfaces within the retromer–Vps5 complex. **a**, Ribbon model of the Vps26 dimer interacting with four membrane-bound Vps5 dimers. Segmented electron density for the lipid bilayer is illustrated. The top view (left) highlights the two-fold symmetry of the Vps26–Vps5 assembly. The cross-section through the model (right) illustrates interactions between Vps26 loops and Vps5 helices. **b**, Ribbon and surface models of Vps5. Surfaces are gradient-coloured by electrostatic potential from red (negative) to blue (positive). **c**, Adjacent Vps5 dimers form tip-to-tip interactions between BAR-domain helices and lateral interactions between PX domains. **d**, Overlay of ribbon model and the electron-density map, showing that the C-terminal α-helix of one Vps5 monomer protrudes towards the Vps26 C-terminal domain. In human VPS26, this is where DMT1-II cargo binds in cooperation with SNX3 (Extended Data Fig. 4c, d). **e**, The apex of the retromer arch viewed looking towards the membrane (left) and from the side (right). It is formed by a homodimeric interaction of Vps35 subunits (interface in blue) on the opposite face to that on which Vps29 is bound. Vps35 residues D694 (D620 in human VPS35) are indicated by arrows. The mutation D620N in human VPS35 causes Parkinson's disease. **f**, Cut-away view showing one leg of the arch. **g**, As in **f**, coloured by sequence conservation from red (low) to blue (high).

© 2018 Springer Nature Limited. All rights reserved.

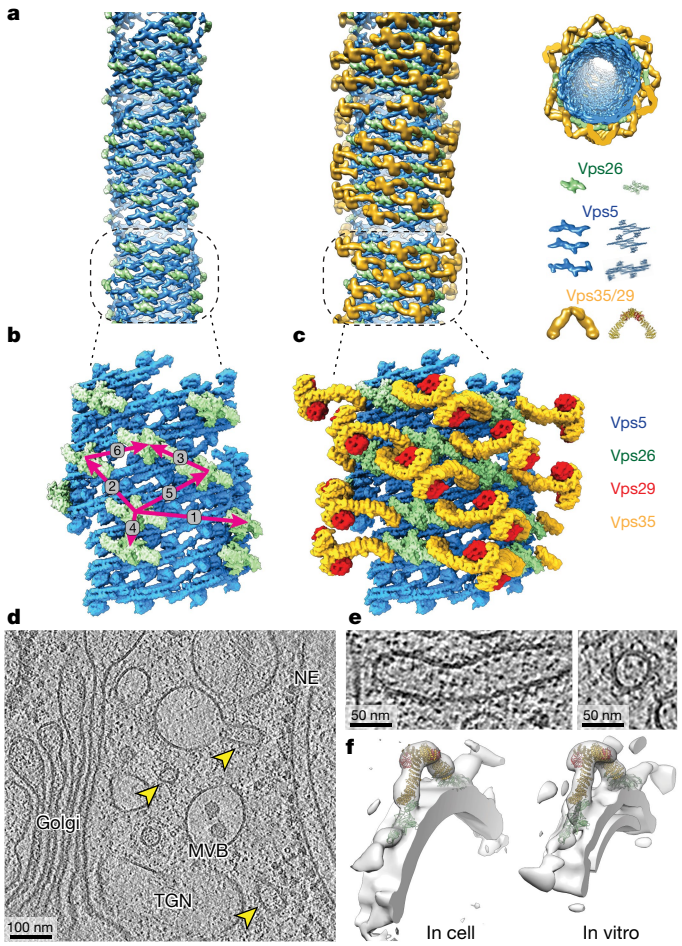


Fig. 3 | Organization of the retromer–Vps5 coat on membrane tubules, and the structure of retromer within the cell. a, A typical retromer–Vps5-coated tubule. Models of the individual elements of retromer–Vps5 have been placed at positions and orientations determined by subtomogram averaging. On the left, the Vps5 and Vps26 layers are shown; the middle and right show the complete coat. On the right is a view along the tube axis of a retromer tubule. Representative models were prepared by segmentation and low-pass filtering of key features, and their respective protein structures are illustrated. **b**, A model of the Vps5–Vps26 layers (corresponding to the dashed box in **a**). Vps26 dimers dock in six relative orientations on the underlying Vps5 array, indicated by magenta arrows (Extended Data Fig. 7e). **c**, A complete model of the retromer coat section shown in **b**. **d**, A slice through one of 12 tomographic reconstructions of a *C. reinhardtii* cell in which retromer-coated membranes were identified (arrowheads). **e**, Magnified views of two of 17 retromer-coated membranes in which arches can be seen. **f**, Density maps filtered to 35 Å from retromer structures determined by subtomogram averaging in situ within the cell and in vitro, fitted with retromer models.

recruitment into retromer–Vps5 tubules could occur through exchange of the SNX proteins that are bound to Vps26 during coat assembly, or through direct binding of the cargo to retromer at different positions in the assembled coat (Extended Data Fig. 4c). Notably, human SNX-BAR proteins were recently found to engage cargo through their PX domains, which suggests that several modes of cargo recognition are possible[21,22].

The Vps35 α-solenoids interact through their N termini with the N-terminal domain of Vps26, and through their C termini with Vps29, as has previously been shown by crystallography[17,18]. Vps35 dimerizes through a conserved interface at the apex of the arch on the opposite surface to Vps29 (Fig. 2e–g). Vps29 is exposed on the outside of the arch apex where it could recruit cytosolic regulatory factors, for example TBC1d5[23] and Varp[24]; this is reminiscent of the recruitment of ArfGAP2 by the exposed γ-COP appendage domain in the COPI coat[25]. The D620N mutation in human VPS35, which causes Parkinson's disease[6,7], lies adjacent to the VPS35 homodimerization interface, suggesting that it may reduce the efficiency of VPS35 dimerization and the subsequent formation of the coat (Fig. 2e).

An extended membrane coat is formed by higher-order oligomerization of the retromer–Vps5 complex. We visualized this higher-order arrangement on individual tubules by placing model retromer and Vps5 complexes at the positions and orientations determined by subtomogram averaging, which revealed a semi-regular membrane coat structure (Fig. 3a). Membrane-associated Vps5 dimers form a pseudo-helical array with limited long-range order, suggesting a plasticity that accommodates varying membrane curvatures. This patchwork Vps5 array provides several potential docking sites, and retromer complexes are positioned unevenly with respect to each other. We analysed the relative positions of Vps26 dimers (Extended Data Fig. 7), finding six different nearest-neighbour pairs corresponding to docking relationships on the Vps5 layer (Fig. 3b, Extended Data Fig. 7e, f). We determined local structures of the retromer coat for each of these preferred arrangements of Vps26 (Extended Data Fig. 7g). Arrangement 1 corresponds to the positioning of Vps26 at each end of a retromer arch, spanning two Vps5 dimers in adjacent rows. Other arrangements result from the binding of Vps26 at alternative positions in nearby rows of the Vps5 array. When Vps26 dimers are docked closely together, the retromer arches can tilt to accommodate closely associated neighbours (Extended Data Fig. 7h). These models were then combined to visualize a section of the retromer coat from a representative tubule (Fig. 3b, c).

We next searched for the presence of protein-coated tubular membranes within a dataset of cryo-electron tomograms of cryo-focused ion-beam-milled *Chlamydomonas reinhardtii* cells[26]. We identified 17 coated tubules emanating from multivesicular bodies, the trans-Golgi network, endosomes and lysosomes (Fig. 3d). In some tubule coats, arch-like features could be directly observed in the tomograms (Fig. 3e). For these cases, we applied reference-free subtomogram averaging to resolve the native, in situ structures of these coats, and found arch-like structures that were essentially identical at the determined resolution to those observed in vitro (Fig. 3f). Although the resolution is insufficient to identify membrane-proximal protein densities, density corresponding to Vps26, Vps29 and Vps35 is clearly visible. This indicates that the arch-like arrangement of retromer observed in vitro represents the architecture present within the cell, and furthermore, that this is conserved between yeast and green algae.

The retromer–SNX-BAR coat observed by cryo-ET (Figs. 1c, 3a–c) is very different from a recently proposed model[17]. Vps29 and Vps35 subunits are exposed on the outside of the coat and are accessible for interaction with peripheral proteins, including regulators of actin and Rab GTPases. Previous cryo-electron microscopy structures of isolated tubules coated with BAR domains have almost exclusively been determined with enforced helical symmetry[13,20,27]. However, the irregular array of Vps5 we observe is more similar to that recently described for Bin1[28], which was also determined without enforcing helical symmetry, and shows that proteins with BAR domains can modulate membrane tubulation without forming a perfect helical scaffold. Our cryo-ET structure suggests that retromer can stabilize the Vps5 array in two ways. First, each Vps26 dimer affixes four Vps5 dimers together. Second, each retromer arch bridges Vps5 proteins that are more than 200 Å apart on the membrane, thereby spanning approximately 60° angular sections of the tubule and enabling retromer to modulate curvature over long distances. The coat architecture is consistent with a model in which cargo recognition is mediated by SNX proteins alone[10,21,22] or together with retromer[17], and in which the primary role of retromer itself is as a structural scaffold that propagates curvature and interacts with regulatory proteins. In general, the retromer–Vps5 architecture shares similarities with other membrane coats, including those formed by COPI, in which α-solenoid subunits provide arch-shaped scaffolds to position membrane-associated Arf1 subunits[29]. This finding supports the proposed ancestral function for extended α-solenoid structures in organizing membrane-bending and cargo-binding modules at the membrane surface.

© 2018 Springer Nature Limited. All rights reserved.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0526-z.

Received: 13 November 2017; Accepted: 8 August 2018;
Published online 17 September 2018.

1. Burd, C. & Cullen, P. J. Retromer: a master conductor of endosome sorting. *Cold Spring Harb. Perspect. Biol.* **6**, a016774 (2014).
2. Seaman, M. N. The retromer complex – endosomal protein recycling and beyond. *J. Cell Sci.* **125**, 4693–4702 (2012).
3. Trousdale, C. & Kim, K. Retromer: structure, function, and roles in mammalian disease. *Eur. J. Cell Biol.* **94**, 513–521 (2015).
4. McMillan, K. J., Korswagen, H. C. & Cullen, P. J. The emerging role of retromer in neuroprotection. *Curr. Opin. Cell Biol.* **47**, 72–82 (2017).
5. Small, S. A. & Petsko, G. A. Retromer in Alzheimer disease, Parkinson disease and other neurological disorders. *Nat. Rev. Neurosci.* **16**, 126–132 (2015).
6. Vilariño-Güell, C. et al. *VPS35* mutations in Parkinson disease. *Am. J. Hum. Genet.* **89**, 162–167 (2011).
7. Zimprich, A. et al. A mutation in *VPS35*, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am. J. Hum. Genet.* **89**, 168–175 (2011).
8. Feng, S. et al. The sorting nexin 3 retromer pathway regulates the cell surface localization and activity of a Wnt-activated polycystin channel complex. *J. Am. Soc. Nephrol.* **28**, 2973–2984 (2017).
9. Bean, B. D., Davey, M. & Conibear, E. Cargo selectivity of yeast sorting nexins. *Traffic* **18**, 110–122 (2017).
10. Gallon, M. et al. A unique PDZ domain and arrestin-like fold interaction reveals mechanistic details of endocytic recycling by SNX27-retromer. *Proc. Natl Acad. Sci. USA* **111**, E3604–E3613 (2014).
11. Zhang, P., Wu, Y., Belenkaya, T. Y. & Lin, X. SNX3 controls Wingless/Wnt secretion through regulating retromer-dependent recycling of Wntless. *Cell Res.* **21**, 1677–1690 (2011).
12. Harterink, M. et al. A SNX3-dependent retromer pathway mediates retrograde transport of the Wnt sorting receptor Wntless and is required for Wnt secretion. *Nat. Cell Biol.* **13**, 914–923 (2011).
13. Mim, C. et al. Structural basis of membrane bending by the N-BAR protein endophilin. *Cell* **149**, 137–145 (2012).
14. Pylypenko, O., Lundmark, R., Rasmuson, E., Carlsson, S. R. & Rak, A. The PX-BAR membrane-remodeling unit of sorting nexin 9. *EMBO J.* **26**, 4788–4800 (2007).
15. Seaman, M. N., McCaffery, J. M. & Emr, S. D. A membrane coat complex essential for endosome-to-Golgi retrograde transport in yeast. *J. Cell Biol.* **142**, 665–681 (1998).
16. Nothwehr, S. F., Bruinsma, P. & Strawn, L. A. Distinct domains within Vps35p mediate the retrieval of two different cargo proteins from the yeast prevacuolar/endosomal compartment. *Mol. Biol. Cell* **10**, 875–890 (1999).
17. Lucas, M. et al. Structural mechanism for cargo recognition by the retromer complex. *Cell* **167**, 1623–1635.e14 (2016).
18. Hierro, A. et al. Functional architecture of the retromer cargo-recognition complex. *Nature* **449**, 1063–1067 (2007).
19. Norwood, S. J. et al. Assembly and solution structure of the core retromer protein complex. *Traffic* **12**, 56–71 (2011).
20. Frost, A. et al. Structural basis of membrane invagination by F-BAR domains. *Cell* **132**, 807–817 (2008).
21. Kvainickas, A. et al. Cargo-selective SNX-BAR proteins mediate retromer trimer independent retrograde transport. *J. Cell Biol.* **216**, 3677–3693 (2017).
22. Simonetti, B., Danson, C. M., Heesom, K. J. & Cullen, P. J. Sequence-dependent cargo recognition by SNX-BARs mediates retromer-independent transport of CI-MPR. *J. Cell Biol.* **216**, 3695–3712 (2017).
23. Jia, D. et al. Structural and mechanistic insights into regulation of the retromer coat by TBC1d5. *Nat. Commun.* **7**, 13305 (2016).
24. Hesketh, G. G. et al. VARP is recruited on to endosomes by direct interaction with retromer, where together they function in export to the cell surface. *Dev. Cell* **29**, 591–606 (2014).
25. Dodonova, S. O. et al. 9Å structure of the COPI coat reveals that the Arf1 GTPase occupies two contrasting molecular environments. *eLife* **6**, e26691 (2017).
26. Bykov, Y. S. et al. The structure of the COPI coat determined within the cell. *eLife* **6**, e32493 (2017).
27. Adam, J., Basnet, N. & Mizuno, N. Structural insights into the cooperative remodeling of membranes by amphiphysin/BIN1. *Sci. Rep.* **5**, 15452 (2015).
28. Daum, B. et al. Supramolecular organization of the human N-BAR domain in shaping the sarcolemma membrane. *J. Struct. Biol.* **194**, 375–382 (2016).
29. Dodonova, S. O. et al. A structure of the COPI coat and the role of coat proteins in membrane vesicle assembly. *Science* **349**, 195–198 (2015).

Acknowledgements We acknowledge support from the University of Queensland Remote Operation Crystallisation Facility, the Centre for Microscopy and Microanalysis (CMM), and the Australian Synchrotron. We thank S. Norwood for assistance and discussion, J. Plitzko and W. Baumeister for enabling in situ cryo-ET, and W. Hagen for supporting data collection. Work in B.M.C.'s laboratory was supported by the Australian Research Council (DP160101743) and National Health and Medical Research Council (NHMRC) (APP1042082; APP1058734). M.S. and B.J.E. were supported by the Max Planck Society. R.D.T. and B.M.C. are supported by NHMRC Senior Research Fellowships (APP1041929; APP1136021). N.L. and D.J.O. are supported by Wellcome Trust grant 207455/Z/17/Z. J.A.G.B., Y.B. and O.K. were supported by the European Molecular Biology Laboratory (EMBL) and the Medical Research Council (MC_UP_1201/16).

Reviewer information *Nature* thanks K. Grünewald and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions O.K., N.L., R.D.T., D.J.O., J.A.G.B. and B.M.C. designed the project. O.K. performed cryo-ET, tomogram reconstruction, subtomogram averaging and model building with assistance from J.A.G.B. O.K. and N.L. performed electron microscopy sample preparation and screening. N.L. performed biochemical characterization and protein preparation, crystal structure determination and assisted with tomogram reconstruction. N.A. assisted N.L. with preliminary electron microscopy sample screening. Y.S.B. performed image analysis for in situ cryo-ET. M.S. and B.D.E. performed in situ cryo-ET. O.K., N.L., J.A.G.B. and B.M.C. prepared the manuscript with input from all authors in the final version.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0526-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0526-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.J.O., J.A.G.B. or B.M.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. All proteins were expressed in *Escherichia coli* BL21(DE3) strain with induction by isopropyl β-D-1-thiogalactopyranoside (Extended Data Table 3). For purification of the retromer complex, Vps29 and Vps35 proteins were co-expressed, whereas Vps26 was expressed separately. Cell pellets from both expressions were combined and lysed by high-pressure homogenization in Buffer R (20 mM HEPES-KOH pH 7.5, 200 mM NaCl). The homogenate was cleared by centrifugation and loaded onto a HisTrap HP (GE Healthcare) resin cartridge followed by elution with 200 mM imidazole in Buffer R. The eluate was loaded onto a Glutathione Sepharose 4B (GE Healthcare) cartridge and the retained protein was eluted by PreScission protease cleavage of the glutathione *S*-transferase (GST) domain. The eluted retromer, devoid of GST-tag, was further purified by gel-filtration chromatography using a Superdex 200 10/300 column (GE Healthcare) in Buffer R with 1 mM tris(2-carboxyethyl)phosphine (TCEP) (Extended Data Fig. 1a).

His-tagged Vps5 dimer was isolated using a HisTrap cartridge followed by gel-filtration chromatography as above (Extended Data Fig. 1b, c). To prepare the retromer–Vps5 complex, the purified Vps5 was mixed with the Vps35–Vps26–Vps29 complex, and the mixture was purified by gel filtration on Superose 6, 10/300 (GE Healthcare) in Buffer R with 1 mM TCEP.

GST-tagged Vps29 was expressed and purified essentially as above but using thrombin (Sigma Aldrich) to cleave the linker between GST and Vps29, and 50 mM Tris-HCl pH 8.0, 200 mM NaCl buffer instead of Buffer R. The cleaved protein was further purified by gel-filtration chromatography on a Superdex 75 10/300 column (GE Healthcare) in 50 mM Tris-HCl pH 8.0, 200 mM NaCl.

Determination of the crystal structure of *C. thermophilum* Vps29. *C. thermophilum* Vps29 is 201 amino acids in length and thrombin cleavage leaves an additional two non-native N-terminal residues (GlySer). Crystals of Vps29 were produced by hanging-drop vapour diffusion using protein at a concentration of 15 mg ml^{-1}. Crystals in space group $P2_1$ grew from a reservoir solution of 8% PEG20000, 8% PEG550MME, 0.2 M calcium acetate and 0.1 M Tris (pH 8.0), and were cryoprotected in 25% glycerol. Crystals were screened at the UQ ROCX diffraction facility on a Rigaku FR-E Superbright generator with Osmic Vari-Max HF optics and Rigaku Saturn 944 CCD detector. Data for structure determination were collected at the Australian Synchrotron MX1 Beamline. Data were integrated with iMOSFLM[30] and scaled with SCALA[31]. The structure was solved by molecular replacement with Phaser[32] using the human VPS29 protein[18] as an input model. The resulting models were rebuilt with Coot[33] and refined with PHENIX[34]. Crystallographic data and structure statistics are provided in Extended Data Table 2.

Liposome pelleting and tubulation. Liposomes composed of POPC/POPE (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine and 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine in a 70:30 molar ratio, Avanti Polar Lipids), Folch brain extract (Sigma) and Folch brain extract supplemented with 3 mol% of dipalmitoyl-phosphatidylinositol-3-phosphate (PI3P) (Echelon Biosciences) were prepared at a liposome concentration of 1 mg ml^{-1} in Buffer R by extrusion through a 0.4 μm polycarbonate filter. For pelleting assays, liposomes at a final concentration of 0.5 mg ml^{-1} and protein at a final concentration of 3 μM were mixed in a final volume of 50 μl, incubated at room temperature for 15 min and then spun at 50,000*g* for 30 min at 20 °C in a Beckman TLA 100 rotor. Supernatant and pellet fractions were analysed by Coomassie-stained SDS–PAGE (Extended Data Fig. 1e). Although the relative stoichiometry of retromer to Vps5 cannot be reliably measured by Coomassie staining, we note that there appears to be a reduced stoichiometry of retromer to Vps5 in the assembled structure. A proportion of retromer may dissociate from the pre-assembled retromer–Vps5 complex as tubule formation proceeds.

For tubulation assays, proteins (2.5 μM) were incubated with Folch liposomes (0.5 mg ml^{-1}) for 2 h at room temperature in Buffer R (Extended Data Fig. 1f).

Sample preparation and image acquisition for cryo-electron microscopy. The retromer–Vps5 liposome-tubulation reaction was mixed with 10-nm gold fiducial markers in identical buffer solution. An aliquot (3 μl) of this mixture was applied to a glow-discharged holey carbon grid (CF-2/1-3C, Protochips), blotted from the back and plunge-frozen in liquid ethane (Leica EM GP automatic plunger).

Imaging was performed on an FEI Titan Krios microscope fitted with a Gatan Quantum 967 LS and a Gatan K2 Summit direct detector operated by Serial-EM software[35]. A total of 71 tomographic series was acquired using a dose-symmetric scheme[36], with tilt range ±60°, 3° angular increment and defoci between −2.5 μm and −6.5 μm. The acquisition magnification was 105,000×, resulting in a calibrated pixel size of 1.35 Å. Tilt images were recorded as ten-frame movies in super-resolution mode at a dose rate of approximately 1.8 e$^-$ Å^{-2} s^{-1} and a total dose per tomogram of around 131 e$^-$ Å^{-2}. Super-resolution 8K frames were aligned, combined and Fourier-cropped to 3,838 × 3,710 pixels using the 'alignframes' command from the IMOD package. Data collection parameters are summarized in Extended Data Table 1.

Tomogram reconstruction and subtomogram averaging. A small number of high-tilt images were discarded (for example, those showing considerable movement, or those in which other objects obscured the field of view). Tilt-series were filtered according to the cumulative radiation dose[37] and aligned on the basis of the gold fiducials using the IMOD package[38]. Contrast transfer function (CTF) correction was performed with the 'ctfphaseflip' IMOD command[39] using defocus values measured by CTFFIND4[40] on non-dose-filtered images. Tomograms were reconstructed from CTF-corrected, aligned stacks using weighted back-projection in IMOD. Tomograms were further binned 2, 4 and 8 times (hereafter called bin2, bin4 and bin8 tomograms) with anti-aliasing, resulting in pixel sizes of 2.7 Å, 5.4 Å and 10.8 Å, respectively.

Subtomogram averaging was performed using MATLAB (MathWorks) functions adapted from the TOM[41], AV3[42] and Dynamo packages[43], essentially as previously described[44]. To define initial subtomogram positions, the central axes of retromer tubules were manually traced in bin4 tomograms and tube radii (measured from the centre of the tube to the middle of the retromer layer) were recorded. The tubular geometry was used to define coordinates and initial rotations of overlapping subvolumes along the tube surface. Rotations were defined to orient subvolumes normal to the membrane and with the in-plane angle perpendicular to the main tubular axis; the subvolume spacing was 12 pixels (21.6 Å).

Reference-free subtomogram averaging was performed independently on 120^3 pixel subtomograms extracted from two bin4 tomograms acquired at −2.5 μm and −5.5 μm defoci. A low-pass filter at 50 Å and a soft-edged cylindrical alignment mask were applied throughout. The alignment in both tomograms converged in five iterations, revealing prominent two-fold symmetrical arch-like structures decorating the membrane layer (Extended Data Fig. 2a). These initial structures were aligned, averaged (Extended Data Fig. 2b) and rotated to place either the arch ('arch' model) or the base between two neighbouring arches ('base' model) in the centre of the reference (Extended Data Fig. 2c). The observed two-fold symmetry was applied, and the reference was low-pass-filtered to 50 Å before using it as a starting reference for alignment of the full dataset.

The full dataset was aligned against the base and arch models in two identical but independent procedures. Initial alignment was performed using 60^3 pixel subvolumes extracted from bin8 tomograms applying a low-pass filter at 50 Å and two-fold symmetry. Upon alignment convergence (Extended Data Fig. 2d), in which two or more subvolumes had converged to similar positions (within 5 pixels distance), only the subvolume giving the highest cross-correlation value to the reference was retained. Retained subvolumes were divided into odd and even datasets, which were processed independently hereafter.

Subsequent iterations of alignment and averaging were performed on subvolumes of 72^3 pixels extracted from bin4 tomograms, and then subvolumes of 144^3 pixels from bin2 tomograms. The step-increment of the angular search was gradually decreased and the low-pass filter was gradually moved towards higher resolution (but set at least three Fourier pixels lower than the frequency corresponding to 0.5 Fourier shell correlation (FSC) between the odd and even datasets). A shell mask that follows the curvature of the membrane layer was applied so that only retromer density was considered during alignment. A population of subvolumes with lowest cross-correlation (around 40%) was removed by imposing a cross-correlation threshold selected manually for each tomogram. The threshold was set to remove subvolumes that were not aligned to the membrane. The resolution for both base and arch alignments reached approximately 13 Å upon convergence (see maps in Extended Data Fig. 2e).

Further alignments were performed using masks focused on three different regions of the arch structure (apex, leg and the membrane-associated region below the apex, Extended Data Fig. 2f, boxes 1, 2 and 3, respectively) and two different regions of the base structure (immediately adjacent to, and slightly above, the membrane; Extended Data Fig. 2f, boxes 4 and 5, respectively) to generate five maps (Extended Data Table 1). Two-fold symmetry was applied except for the alignment focused on the leg, where two subvolumes were extracted from each arch using dynamo_subboxing_table (Dynamo), and symmetry was not applied.

Final converged averages were formed by 37,360 asymmetric units for arch-derived and 32,074 asymmetric units for base-derived maps. The measured overall resolutions for each average ranged from 8.8 Å to 9.7 Å (Extended Data Fig. 3a). Local resolution maps are provided in Extended Data Fig. 3b. Final maps were reweighted by division by their summed CTFs and sharpened by empirically determined B factors of between 1,300 and 1,600.

Homology modelling. The deposited atomic model with the highest sequence similarity to the *C. thermophilum* Vps5 BAR domain is that of human SNX1 (Protein Data Bank (PDB) ID: 4FSZ). However, the SNX1 BAR dimer in 4FSZ is highly curved and cannot be docked into the Vps5 BAR densities in our reconstruction. By contrast, the two deposited structures of SNX-BAR domains, SNX9

© 2018 Springer Nature Limited. All rights reserved.

(PDB ID: 2RAJ) and SNX33 (PDB ID: 4AKV), correspond well with the shape of our experimentally determined Vps5 densities. To enable SNX1 to be used as a template for structural homology modelling, we adapted the curvature of SNX1 to match that of SNX9. This was done by flexibly fitting the SNX1 BAR domain into a density map simulated from the SNX9 structure using VMD and MDFF software (within NAMD)[45]. Changes in the secondary structure were prohibited throughout the simulation. The resulting 'curvature adjusted' SNX1 BAR model was then used to build a homology model of the Vps5 BAR domain using the i-TASSER online service[46]. The Vps5 PX domain was modelled on the SNX7 PX domain (PDB ID: 3IQ2) with i-TASSER. A complete Vps5 PX-BAR dimer was modelled by aligning the models of Vps5 PX and BAR domains to the SNX9 template as rigid bodies using the match-maker tool in Chimera.

No structures of the full-length retromer complex are available. Two partial homology models were generated using crystal structures of an N-terminal fragment of human VPS35 bound to VPS26 (PDB ID: 5F0J) and a C-terminal fragment of human VPS35 bound to VPS29 (PDB ID: 2R17). The VPS26–VPS35(N) complex and VPS29–VPS35(C) complex include residues 14–470 and 476–780, respectively, together covering nearly the entire human VPS35 sequence (796 residues in total). These two models were used as templates for modelling *C. thermophilum* Vps26–Vps35(N) and Vps29–Vps35(C), respectively, using the 'hetero-project' pipeline on the Swiss-Model web-server[47]. In the latter model, the modelled Vps29 was then replaced by our X-ray structure of *C. thermophilum* Vps29 (PDB ID: 1W8M).

Model building. UCSF Chimera[48] and ChimeraX[49] were used for rigid-body fitting and visualization. The relative orientations of the locally aligned maps were determined by sequential alignment of the maps using their overlapping regions using the 'fit' command preceded by rough manual placement. The obtained composite was used to visualize the relative positions of structural models built in individual locally aligned maps and to transfer models between the maps.

Vps26–Vps35(N), Vps29–Vps35(C) and Vps5 dimer homology models were fit into locally aligned maps using the Chimera fit command with random initial placement. Unique docking positions were identified displaying the highest cross-correlation and coverage scores (Extended Data Fig. 5a–c).

The rigid-body fitting data showed excellent correspondence with the observed density. For Vps5 and the central region of Vps35, some small differences in the positions of helices were observed between the fitted model and the density in the electron micrograph. Vps5 and residues 122–816 of Vps35 were therefore flexibly fit into the density maps with the highest local resolution using the MDFF software within NAMD[45], maintaining all secondary structure. A long yeast-specific loop (305–387) and a long linker between the N- and C portions of Vps35 (537–559) were excluded from the Vps35 model. An unoccupied density (indicated by the arrowhead in Extended Data Fig. 3b) may correspond to the putative helical region (340–370) in the 305–387 loop of Vps35.

We note that Vps5 dimers that contact Vps26 via their C-terminal helix ('stabilized dimers') are better resolved than dimers with free C termini (Extended Data Fig. 3b). In the final model, the 'stabilized' Vps5 dimer that was fit into the map with the highest local resolution was copied to all other positions in which 'stabilized' dimers were present and was fitted locally as a rigid body. For 'unstabilized' Vps5 dimers, the fit into the map with highest local resolution was copied to another equivalent position.

Arrangement of the retromer coat. The arrangement of the retromer coat on individual tubules was visualized by placing models of protein components of the coat at the coordinates and rotations to which subtomograms converged during alignment (see example in Fig. 3a). Simplified models were prepared by segmenting densities of interest from the corresponding Gaussian-smoothed subtomogram averages using UCSF Chimera.

To assess the long-range order of the retromer coat, we determined the relative orientation of neighbouring retromer subunits. To do this, we took the positions and rotational orientations of all subtomograms in the final alignment, and plotted the relative positions of all neighbouring subtomograms (those within 416 Å). We performed this operation for subtomograms aligned on Vps26 and for subtomograms aligned on the apex of the retromer arch. In this way, the positions of 16,037 and 18,680 sets, respectively, of distance-selected neighbours were plotted to generate a 3D density plot (Extended Data Fig. 7a, b). To simplify visualization, we flattened and projected the 3D density plot to give a 2D density plot (Extended Data Fig. 7c, d). The presence of peaks in the density plot shows that neighbours are found in preferred relative positions. The membrane-associated layer of the coat (represented by Vps26 dimers) shows more long-range order than do the retromer arches, probably owing to relative structural flexibility in the arch (compare Extended Data Fig. 7c and d).

Analysing the 3D density plot for Vps26 neighbours reveals six positions in which the nearest-neighbour Vps26 is located. Subtomograms were divided into groups according to whether a neighbour was identified in each of the six peaks (numbered in Extended Data Fig. 7e). Peaks related by the two-fold symmetry of Vps26 are equivalent and were therefore included in the same groups. Each of the six groups of subtomograms was then separately averaged (at bin4) to generate six local density maps of the retromer coat (Extended Data Fig. 7g). Each of the structures therefore shows the conformation of retromer when a nearest-neighbour is present in one of the six preferred positions (Fig. 3b).

Models shown in Fig. 3a were prepared by segmentation and low-pass filtering; their respective protein structures are illustrated in the box. The models shown in Fig. 3b, c were built by fitting retromer components into the six density maps for each relation (Extended Data Fig. 7g), and combining the fitted structures in the relative arrangements observed in Fig. 3a.

Determination of the retromer coat structure within *C. reinhardtii*. In a previously described study, *C. reinhardtii* cells were vitrified, and thin lamellae containing regions of the cell interior were prepared by cryo-focused ion-beam milling[50,51] before imaging by cryo-electron tomography[26,52]. We reanalysed these tomograms, identifying 17 coated tubular membranes in 12 tomograms. The coat on these tubes appeared distinct from previously observed COPI, COPII and clathrin coats. For nine tubes, we could determine the morphology of the compartment from which they emanated: four budded from multivesicular bodies, four from trans-Golgi network compartments (rounded translucent cisternae without luminal vesicles), and one from a lysosome-like compartment. In tomographic slices of some of the tubes positioned perpendicular to the XY plane, we observed that the coat was formed of arch-like densities. We performed reference-free subtomogram averaging to determine the structure of the coat, essentially as previously described[26], using a modified wedge mask that accounts for the amplitudes of the determined CTF and applied exposure filters at each tilt[53]. We extracted subtomograms from binned tomograms (pixel size 1.36 nm) normal to the tube surface with a spacing of 8 pixels (11 nm) and box size of 64 pixels. Subtomogram positions were randomized by 4 pixels along the tube axis to avoid any artefacts from periodic picking. The initial average was created and subtomograms were aligned to it with only *z*-axis shifts, followed by 12 iterations of alignment with a soft cylindrical mask of 16-nm diameter and 37-nm height, with a low-pass filter set to 60 Å. The resulting average displayed two arches. The centermost arch was shifted to the centre of the box and used as a reference for four additional iterations of alignment with two-fold symmetry applied. Overlapping subtomograms were removed with the distance threshold of 8 pixels (11 nm), and the dataset was split in two halves from which the averages were generated. The resolution was estimated using mask-corrected FSC between these references, and was 37 Å at 0.5 cross-correlation, extending considerably beyond the 60 Å low-pass filter used during alignment.

Data availability

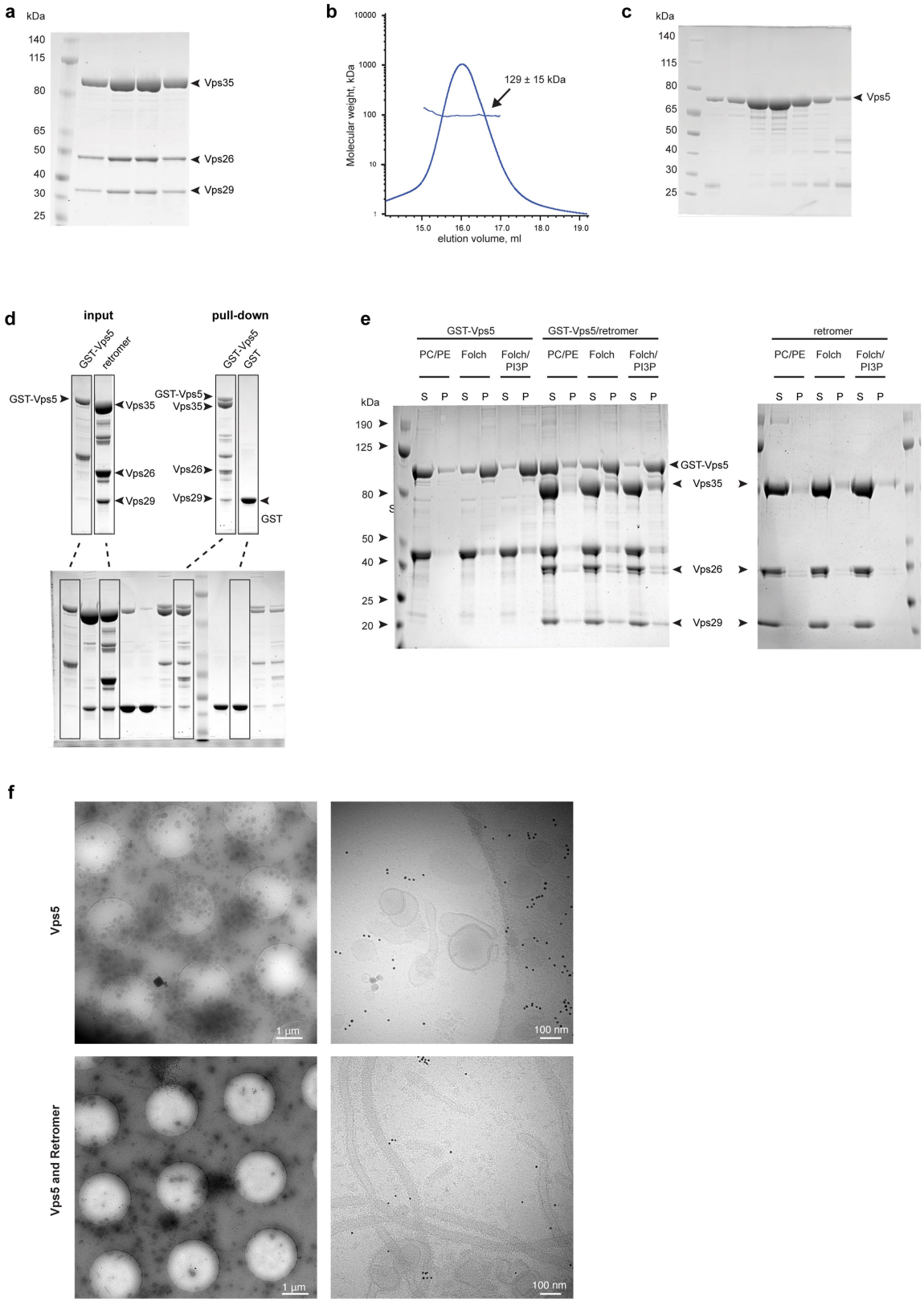
Cryo-ET structures and representative tomograms have been deposited in the Electron Microscopy Data Bank (EMDB) under accession codes EMD-0154, EMD-0155, EMD-0156, EMD-0157, EMD-0158, EMD-0159, EMD-0160, EMD-0161, EMD-0162 and EMD-0163, and the molecular model in the Protein Data Bank (PDB) under accession code 6H7W. Coordinates and structure factors for Vps29 are in the PDB under accession code 5W8M, with raw data available at https://espace.library.uq.edu.au/view/UQ:682629. Gel source data can be found in Supplementary Fig. 1.

30. Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67**, 271–281 (2011).
31. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62**, 72–82 (2006).
32. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
33. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
34. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
35. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
36. Hagen, W. J. H., Wan, W. & Briggs, J. A. G. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. *J. Struct. Biol.* **197**, 191–198 (2017).
37. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).
38. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).
39. Xiong, Q., Morphew, M. K., Schwartz, C. L., Hoenger, A. H. & Mastronarde, D. N. CTF determination and correction for low dose tomographic tilt series. *J. Struct. Biol.* **168**, 378–387 (2009).
40. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
41. Förster, F., Medalia, O., Zauberman, N., Baumeister, W. & Fass, D. Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. *Proc. Natl Acad. Sci. USA* **102**, 4729–4734 (2005).

© 2018 Springer Nature Limited. All rights reserved.

42. Nickell, S. et al. TOM software toolbox: acquisition and analysis for electron tomography. *J. Struct. Biol.* **149**, 227–234 (2005).
43. Castaño-Díez, D., Kudryashev, M., Arheit, M. & Stahlberg, H. Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. *J. Struct. Biol.* **178**, 139–151 (2012).
44. Schur, F. K. et al. An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science* **353**, 506–508 (2016).
45. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
46. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
47. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–W258 (2014).
48. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
49. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
50. Schaffer, M. et al. Cryo-focused ion beam sample preparation for imaging vitreous cells by cryo-electron tomography. *Bio Protoc.* **5**, e1575 (2015).
51. Schaffer, M. et al. Optimized cryo-focused ion beam sample preparation aimed at in situ structural studies of membrane proteins. *J. Struct. Biol.* **197**, 73–82 (2017).
52. Albert, S. et al. Proteasomes tether to two distinct sites at the nuclear pore complex. *Proc. Natl Acad. Sci. USA* **114**, 13726–13731 (2017).
53. Wan, W. et al. Structure and assembly of the Ebola virus nucleocapsid. *Nature* **551**, 394–397 (2017).
54. Bärlocher, K. et al. Structural insights into *Legionella* RidL-Vps29 retromer subunit interaction reveal displacement of the regulator TBC1D5. *Nat. Commun.* **8**, 1543 (2017).
55. Romano-Moreno, M. et al. Molecular mechanism for the subversion of the retromer coat by the *Legionella* effector RidL. *Proc. Natl Acad. Sci. USA* **114**, E11151–E11160 (2017).
56. Harrison, M. S. et al. A mechanism for retromer endosomal coat complex assembly with cargo. *Proc. Natl Acad. Sci. USA* **111**, 267–272 (2014).
57. Liu, T. T., Gomez, T. S., Sackey, B. K., Billadeau, D. D. & Burd, C. G. Rab GTPase regulation of retromer-mediated cargo export during endosome maturation. *Mol. Biol. Cell* **23**, 2505–2515 (2012).
58. Purushothaman, L. K., Arlt, H., Kuhlee, A., Raunser, S. & Ungermann, C. Retromer-driven membrane tubulation separates endosomal recycling from Rab7/Ypt7-dependent fusion. *Mol. Biol. Cell* **28**, 783–791 (2017).
59. Arlt, H., Reggiori, F. & Ungermann, C. Retromer and the dynamin Vps1 cooperate in the retrieval of transmembrane proteins from vacuoles. *J. Cell Sci.* **128**, 645–655 (2015).
60. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890 (1988).
61. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

© 2018 Springer Nature Limited. All rights reserved.

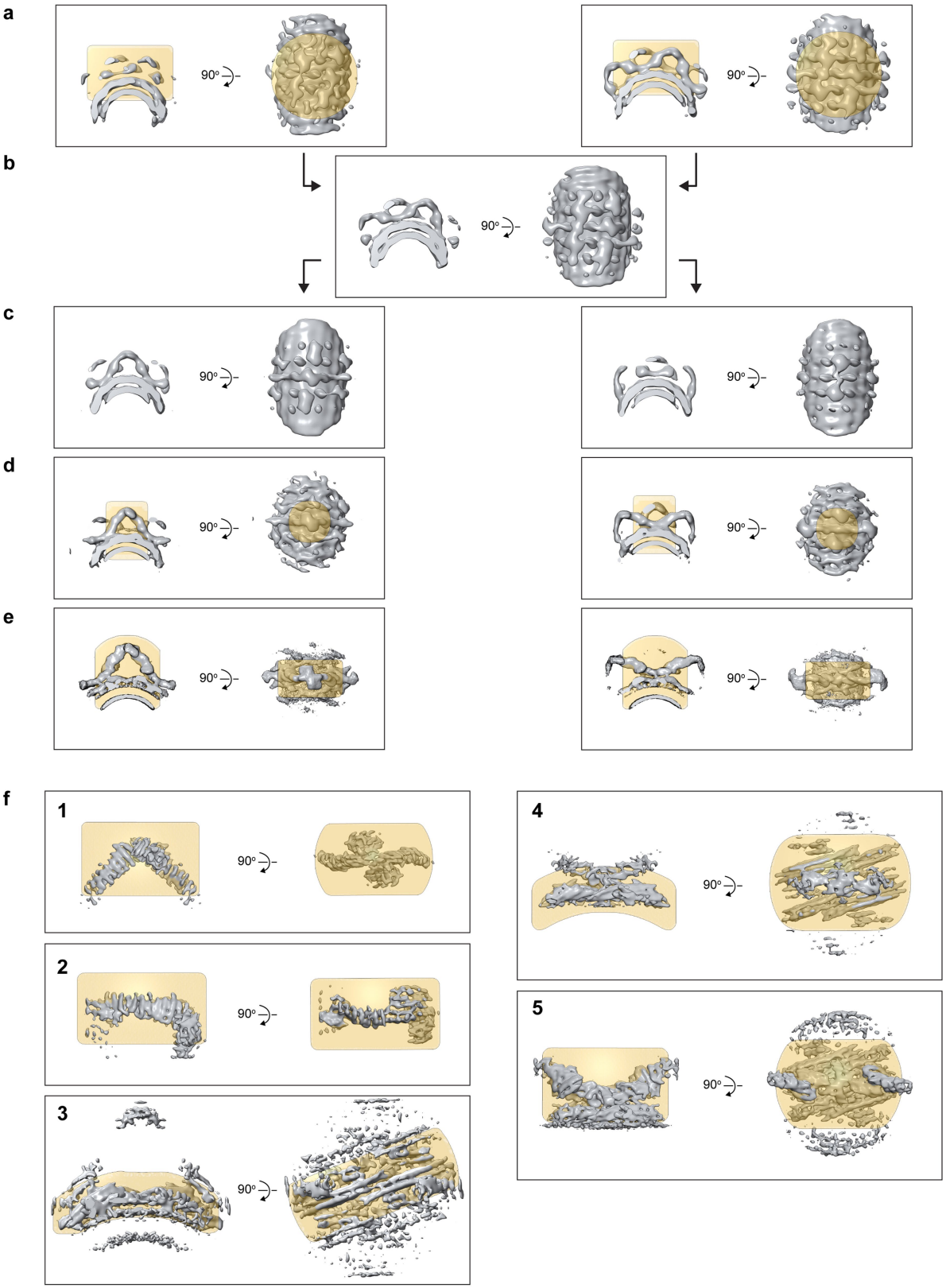


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

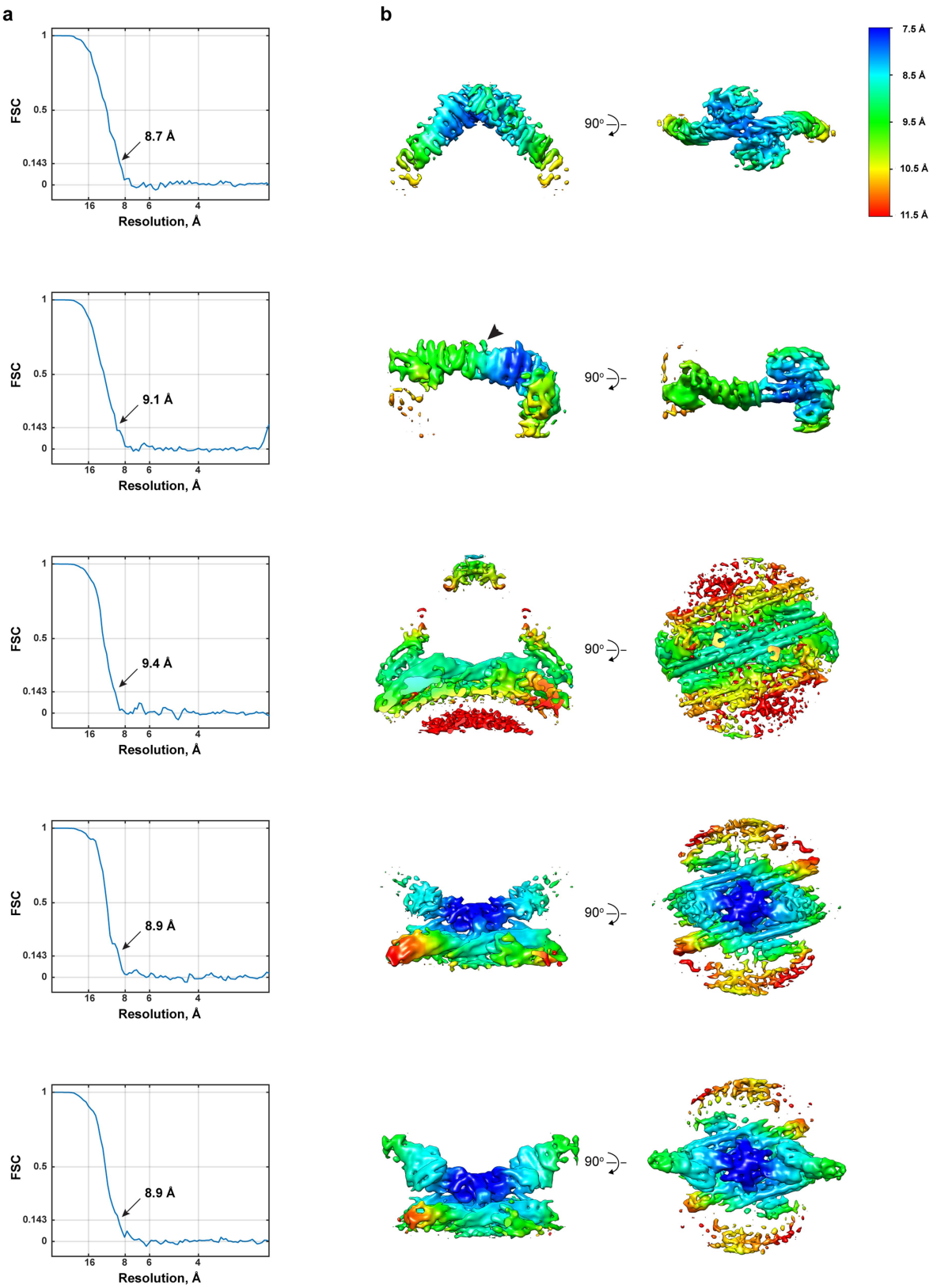
Extended Data Fig. 1 | The retromer–Vps5 complex in solution and binding to membranes. a, Retromer forms a stable complex in solution. Fractions containing retromer (Vps35, Vps26 and Vps29) after gel-filtration on a Superdex 200 column analysed by Coomassie-stained SDS–PAGE. **b**, Gel-filtration profile of Vps5 and multiangle laser light scattering analysis of molecular weight. Mean molecular weight and standard deviation from three independent gel-filtration experiments are shown. The expected molecular weight of Vps5 monomer is 67 kDa, so the observed molecular weight of 129 kDa indicates formation of a homodimer. **c**, A Coomassie-stained gel from SDS–PAGE of Vps5 fractions from **b**, indicating the formation of a homodimer. **d**, Vps5 binds to retromer in solution. SDS–PAGE gels of GST–Vps5 and of retromer are shown under 'input'. Retromer was incubated with GST-tagged Vps5 or GST baits, and the resultant complex was isolated on glutathione sepharose beads ('pull-down'). At the bottom is the intact PAGE gel from which the input and pull-down lanes were extracted. **e**, Membrane recruitment of retromer is dependent on Vps5. GST–Vps5 alone, GST–Vps5 with retromer complex and retromer complex alone were incubated with liposomes and pelleted to isolate the liposome-bound protein fraction. The supernatant (S) and pelleted fraction (P) were compared with Coomassie-stained SDS–PAGE gels. POPC/POPE liposomes were used as a negative control. Vps5 is efficiently pelleted by Folch brain extract liposomes, and the introduction of PI3P does not increase the amount of pelleted protein. The retromer complex shows no membrane association on its own, but is recruited to Folch membranes and membranes containing both Folch extract and PI3P when it interacts with Vps5. **f**, Retromer promotes tubule formation by Vps5. Characteristic cryo-electron microscopy images at medium (left) and high (right) magnification of Folch liposomes incubated either with Vps5 alone (top) or in the presence of the retromer complex (bottom). Data shown in all panels are representative of at least three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



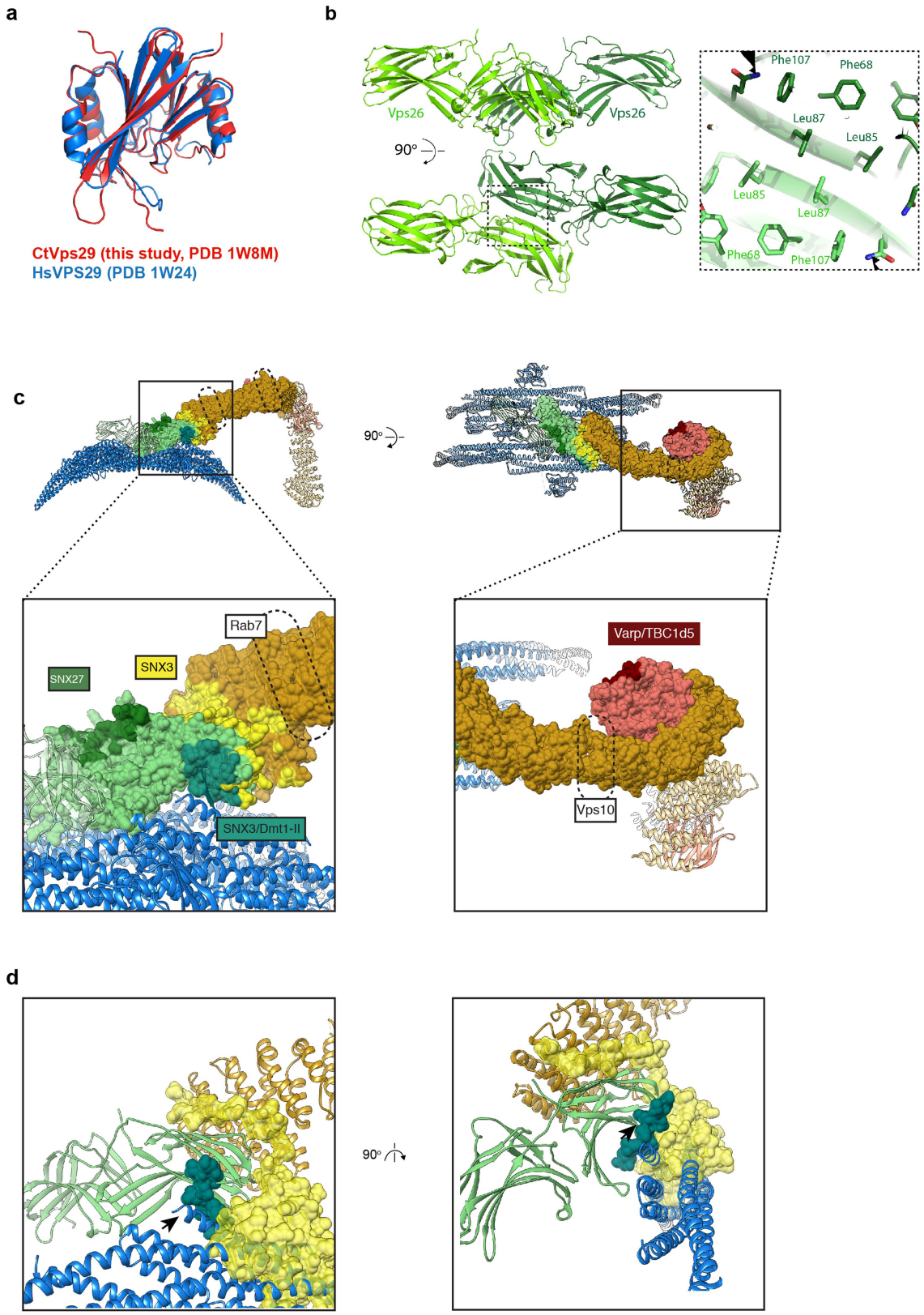
Extended Data Fig. 2 | Overview of the subtomogram-averaging procedure. Stages in the subtomogram averaging procedure are shown from top to bottom. Key steps are illustrated by average volumes (grey) overlaid with the corresponding alignment mask (gold). Alignment masks are shown at a 0.5 value threshold. Volumes in **a–c** are low-pass-filtered to 50 Å. **a**, The final iteration of the reference-free subtomogram-averaging procedure independently conducted in bin4 tomograms that were acquired at −2.5 mm (left) and −5.5 mm (right) defoci, filtered to 50 Å resolution. **b**, The average of the references shown in **a**. **c**, The volume from **b** was rotated to place either the apex of the arch (left) or the base of the arch (right) in the box centre, two-fold symmetrized, and filtered to 50 Å resolution. These two volumes were used as starting references for further alignments. **d**, The references after alignment at bin8. **e**, The references after alignment at bin2. After subtomogram-averaging convergence in bin2, focused alignment was conducted on individual structural features. **f**, The final maps for alignments focused on the regions within the gold alignment masks (boxes 1–3, arch structure; boxes 4 and 5, base structure).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Cryo-ET map and reconstruction resolution. **a**, Mask-corrected FSC curves for each of the final focused maps shown in Extended Data Fig. 2f. The overall resolution at the 0.143 criterion is marked. **b**, Sharpened maps coloured by local resolution according to the indicated colour map determined by FSC within a moving local mask. Arrowheads indicate an unassigned density, which may correspond to a helical element in loop 305–387 of Vps35.

© 2018 Springer Nature Limited. All rights reserved.

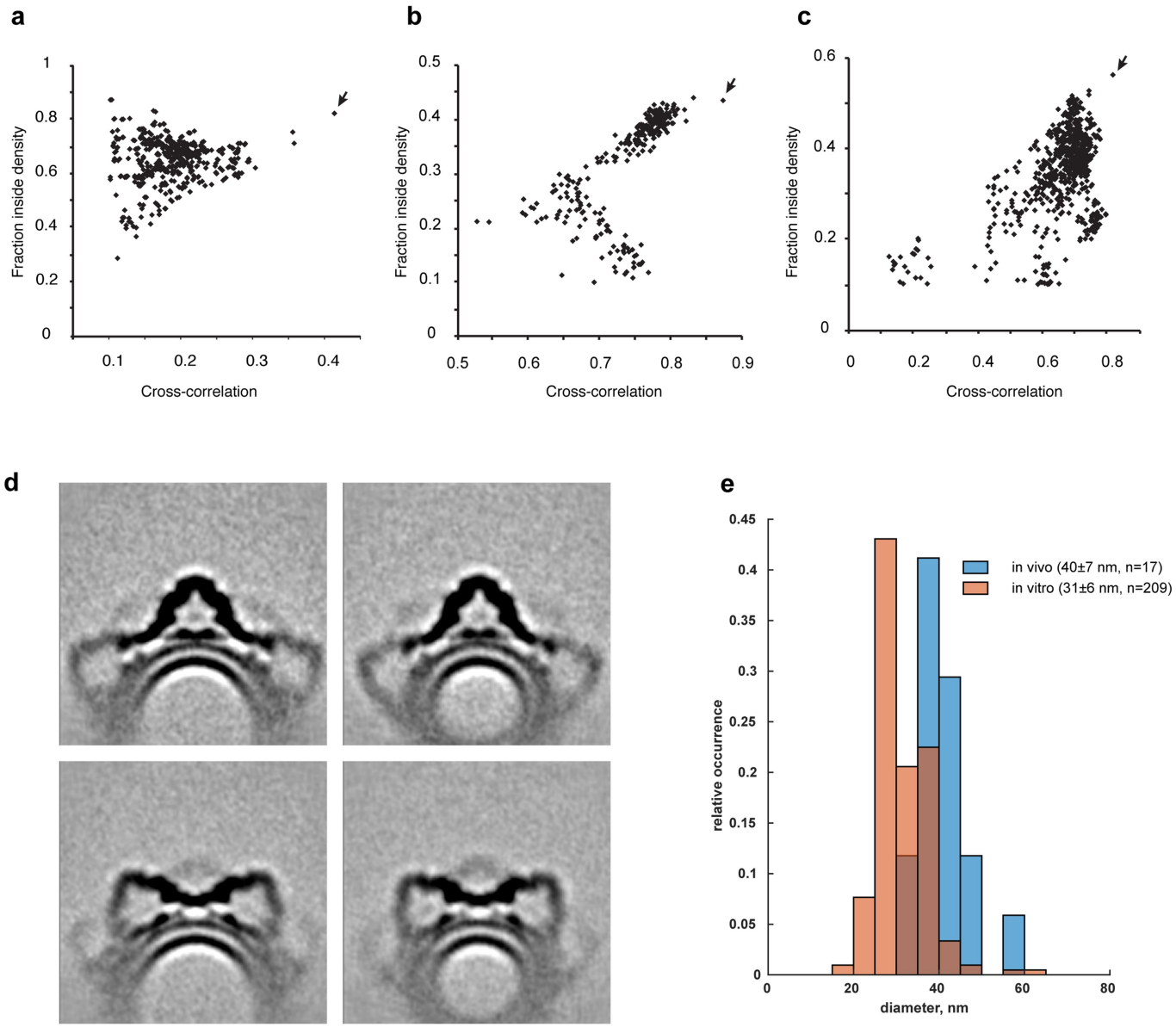


Extended Data Fig. 4 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

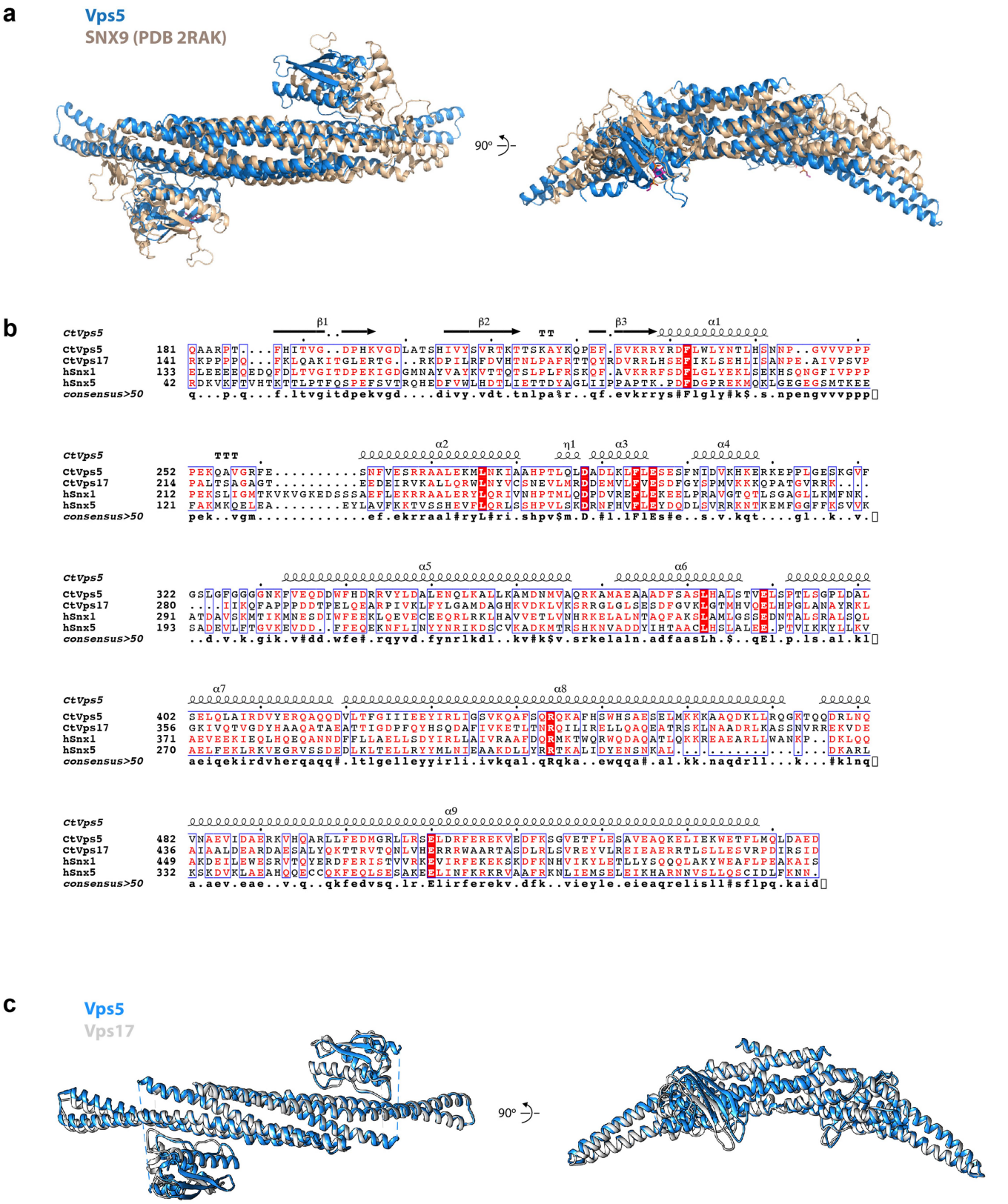
Extended Data Fig. 4 | Crystal structure of *C. thermophilum* Vps29, and comparison of cryo-ET structures with previous crystal structures. **a**, Crystal structure of *C. thermophilum* Vps29 (red) overlaid with the crystal structure of human VPS29 (blue)[18]. Crystallographic structure determination statistics are given in Extended Data Table 2. **b**, The fitted Vps26 dimer model with monomers coloured in dark green and light green. The homodimeric interface is formed by β-sheet extension of two N-terminal β-sandwich domains. The positions of the docked Vps26 models suggest the formation of an extended hydrophobic core between subunits. Close-up images of fitted Vps26 subunits highlight the extended hydrophobic core. **c**, Surface representation of the Vps26–Vps35–Vps29 trimer mapped with binding regions for retromer effectors. Neighbouring Vps5 and retromer proteins in the assembled array are shown as ribbons. Retromer components are coloured as in Fig. 1. The bottom boxes show higher-magnification views of the indicated regions of the overviews. Binding sites observed in structural data are coloured according to the colour of the corresponding label; dashed lines indicate binding regions identified in biochemical assays. The binding interfaces of human SNX27, SNX3, SNX3–DMT1-II, VARP–TBC1D5 were modelled using coordinates with PDB accession numbers 4P2A, 5F0L and 5GTU, respectively; the DMT1-II cargo peptide is shown as a ribbon. The SNX3–DMT1-II binding site overlaps with that of Vps5. The binding site of the SNX27 PDZ domain on VPS26[10] is accessible, although owing to a lack of structural information regarding full-length SNX27 it is unclear whether this binding is consistent with simultaneous membrane binding by the SNX27 PX domains. The regulatory factors VARP and TBC1D5 share a binding interface on VPS29 that is exposed towards the outer extremity of the coat. This site in human VPS29 is also hijacked by the RidL protein from the pathogen *Legionella pneumophila*[54,55]. However, as for SNX27, full-length structures of VARP, TBC1D5 and RidL are not available so we cannot be sure how they will be arranged in the fully assembled array. Rab7 has been speculated to contribute to membrane recruitment of retromer by binding to VPS35 in the region indicated by the dashed line[56,57]. The deletion of this helical region (helix 6 in *Saccharomyces cerevisiae*) resulted in loss of interaction with Ypt7[57] (the Rab7 homologue in yeast). It has recently been shown that retromer binding to the PX-BAR complex displaces Rab7 during the formation of tubules[58,59]. The VPS10 'binding site' (dashed line) indicates a region where point mutations affect VPS10 recycling[16]; however, no biochemical interactions between Vps10 and retromer have been shown, and our efforts at detecting a physical interaction between Vps35 and the cytosolic domains of Vps10 have not shown any direct binding. **d**, SNX3 and DMT1-II (transparent yellow and dark green surfaces respectively), as bound to VPS35–VPS26 from PDB 5F0L[17], overlaid with our retromer–Vps5 complex structure (ribbons, coloured as above) demonstrating a steric clash between Vps5 BAR and SNX3 PX domains. Note that the C-terminal helix of Vps5 BAR (arrowheads) clashes with the DMT1-II cargo peptide density. The left box shows the same view as that above in **c**; the right panel shows the model rotated by 90° around the vertical axis to provide the view along the long axis of the BAR domain.

© 2018 Springer Nature Limited. All rights reserved.



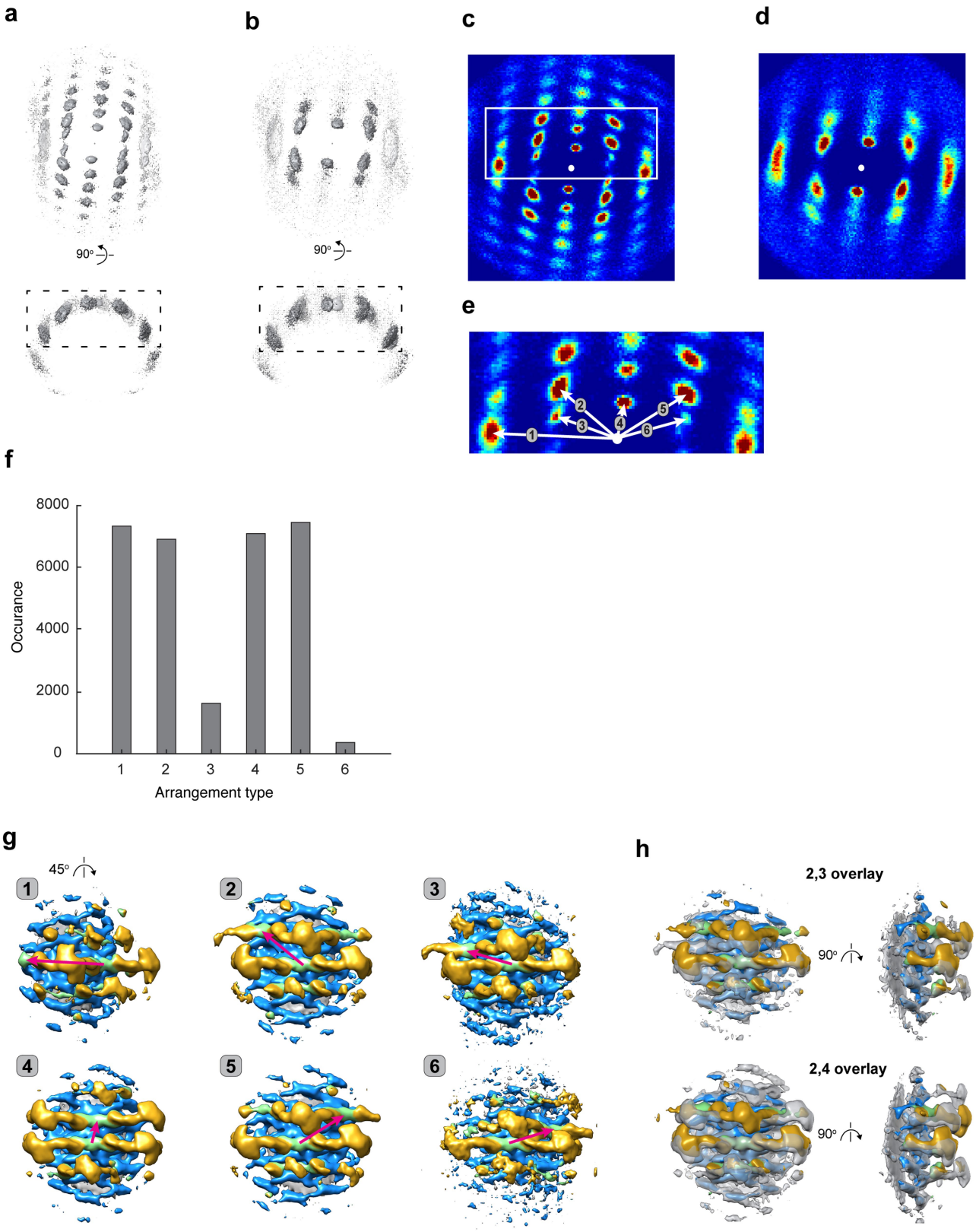
Extended Data Fig. 5 | Global rigid-body docking of Vps5, Vps26–Vps35(N) and Vps35(C)–Vps29, and adaptation of the retromer coat to different membrane curvatures. a–**c**, Fitting of structures to electron density maps was performed from 10,000 random initial placements of atomic models using the 'fit' command in Chimera. **a**–**c**, The cross-correlation between the model and the electron microscopy map is plotted against the fraction of the structural model within the electron-microscopy density threshold for: the global fit of the Vps5 dimer into the membrane-associated BAR domain density under the arch (**a**) (map f3 in Extended Data Fig. 2); the global fit of Vps26–Vps35(N) into the base of the arch (**b**) (map f5 in Extended Data Fig. 2); the global fit of Vps29/Vps35(C) into the apex of the arch (**c**) (map f1 in Extended Data Fig. 2). Arrows indicate the high-scoring rigid-body fits that were used as starting points for flexible fitting. **d**, **e**, For a subset of about 50% of the data we calculated tube centroids by spline fitting, and determined local membrane-curvature as the inverse of the distance from the subtomogram to the tube centroid. **d**, Slices through averages of 20% of the subtomograms from the dataset with lowest (left) and highest (right) membrane curvature, focused on the arch (top) or Vps26 dimer (bottom). See also the animation in Supplementary Video 5. **e**, The distribution of membrane curvatures of retromer tubules in vivo and in vitro. The lumenal diameters of each tube were measured manually, from which the mean and standard deviation were calculated.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Comparison of yeast PX-BAR proteins Vps5 and Vps17. a, Overlay of the Vps5 heterodimer model after flexible fitting into the cryo-ET structure (blue) with the human SNX9 PX-BAR domains (beige; the bound PI3P head group is also shown in magenta in stick representation). The PX and BAR domains in Vps5 adopt a very similar architecture to the SNX9 protein, but there are variations in the angle between the BAR domains, and in the orientations of the lateral PX domains. The second and third α-helices of the Vps5 BAR domain are also longer than those of SNX9. **b**, Sequence alignment of the *C. thermophilum* PX-BAR proteins CtVps5 and CtVps17 and the human PX-BAR proteins SNX1 and SNX5). The secondary structure of CtVps5 is indicated above the sequences. The sequence alignment and its representation were prepared in MultAlign[60] and ESPript 3.0[61]. **c**, Overlay of ribbon models of CtVps5 (blue) and CtVps17 (grey). The structure of CtVps17 was modelled using CtVps5 as a template (SWISS-MODEL).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 7 | Arrangement of the retromer coat on membrane tubules by cryo-ET. a, **b**, 3D plots that visualize the relative positions of neighbouring Vps26 dimers (**a**) or neighbouring Vps35–Vps29 arches (**b**) (see Supplementary Information for details). The isosurface for visualization is set at 8σ. **c**, **d**, Flattened cylindrical projections through the boxed regions in volumes **a**, **b**, respectively. The blue (low) to red (high) gradient colouring is proportional to pixel values. The white circle shows the position of the central Vps26 dimer or arch. **e**, A close-up view of the boxed region in **c** with arrows indicating the position and the identity of neighbours corresponding to each of six nearest-neighbour relative arrangements between Vps26 dimers. **f**, Bar plot of the occurrence of arrangements from **e** for 15,795 analysed Vps26 dimers. The arrangements numbered 3 and 6, in which Vps26 dimers are very closely packed, are less frequent than other arrangements. Models of these relative arrangements are shown in Fig. 3b, c. **g**, Density maps are shown for the local retromer structure for each of the six different relative Vps26 arrangements. Numbering corresponds to arrangements shown in Fig. 3b. Maps are radially coloured in grey, blue, green and gold for the membrane, Vps5, Vps26 and Vps35–Vps29 layers, respectively. **h**, Overlay of density maps for arrangements 2 and 3, and 2 and 4, showing that in some arrangements, the Vps35 arch can tilt relative to the tubule to accommodate nearby arches. The average of arrangement 2 is coloured as above, whereas the averages of arrangement 3 and 4 are coloured transparent grey.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Cryo-ET data collection and processing parameters

	1	2	3	4	5	In cell
EMDB entry	**0156**	**0157**	**0158**	**0159**	**0160**	**0161**
Voltage (kV)	300					
Detector	Gatan K2					
Energy filter slit width (eV)	20					
Electron exposure (e/ Å^2)	~130					~100
Defocus range (μm)	-2.5 to -6.5					-4.5 to -6.5
Tilt range (min/max, step)	-60°/+60, 3°					-60°/+60, 2°
Tilt scheme	dose-symmetrical (Hagen scheme)					bi-directional
Tomogram used/acquired (no.)	71/73					12/60
Pixel size (Å)	1.35					3.42
Tubules (no.)	209					17
Symmetry imposed	C2	none	C2	C2	C2	C2
Final subtomograms (no.)	18680	37250	18680	16037	16037	553
Map resolution (Å)	8.7	9.1	9.4	8.9	8.9	37
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.5
Map resolution range (Å)	7.6-22.1	7.5-20.8	7.1-19.4	6.7-19.5	7.2-19.3	~

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | X-ray crystallography data collection and refinement statistics for *C. thermophilum* Vps29

	Ct Vps29
Data collection	
Space group	$P2_1$
Cell dimensions	
a, *b*, *c* (Å)	65.5, 106.7, 88.1
α, β, γ (°)	90, 95.5, 90
Resolution (Å)	48.8-1.52 (1.55-1.52)*
R_{merge}	0.094 (0.707)
R_{meas}	0.110 (0.0.831)
R_{pim}	0.057 (0.431)
$I / \sigma I$	9.9 (2.1)
Completeness (%)	99.8 (97.7)
Redundancy	7.5 (7.2)
$CC_{1/2}$	0.997 (0.681)
Refinement	
Resolution (Å)	45.59-1.52 (1.56-1.52)
No. reflections/No. R_{free}	184620 (12809) / 2000 (140)
R_{work}/R_{free}	0.155 (0.228) / 0.176 (0.246)
No. atoms	
Protein	8939
Water	1304
B-factors	
Protein	22.8
Water	29.6
R.m.s. deviations	
Bond lengths (Å)	0.0097
Bond angles (°)	1.30

*Values in parentheses are for the highest-resolution shell.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 3 | *E. coli* expression constructs for *C. thermophilum* proteins used in this study

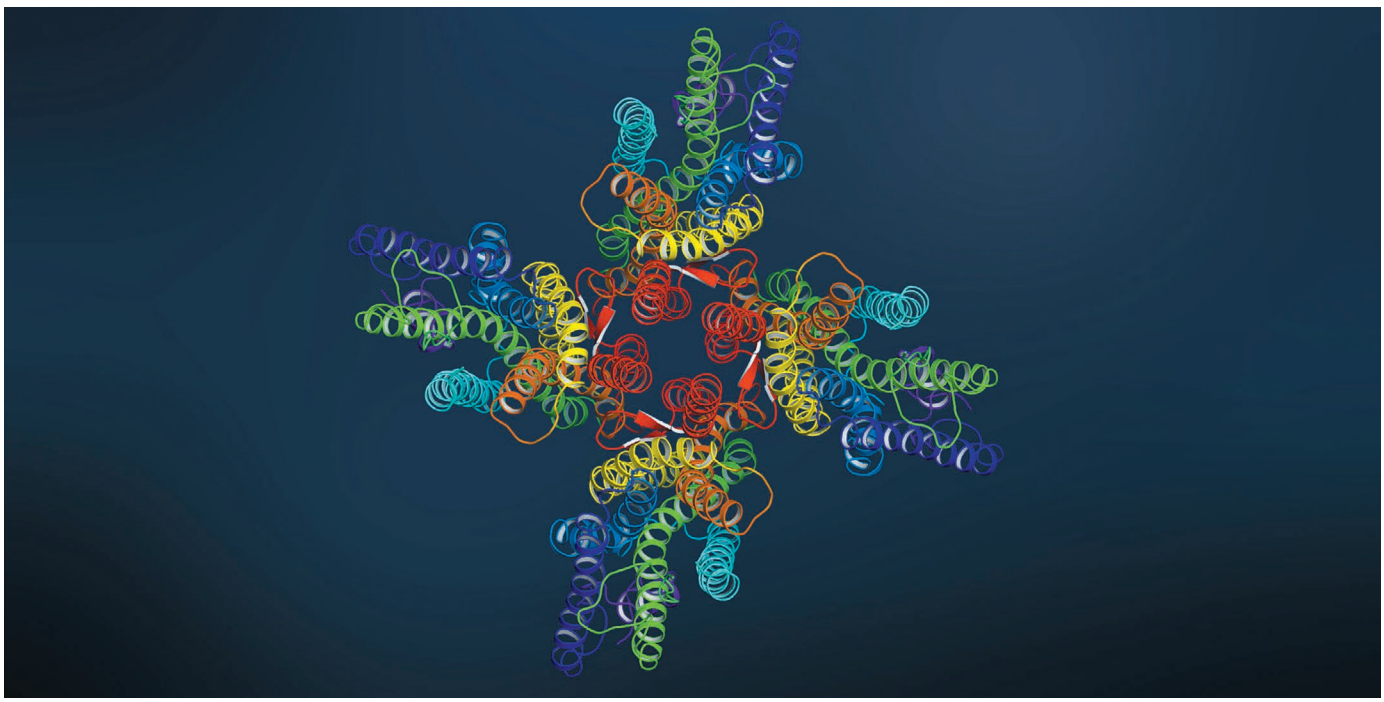
Name of construct	Method of cloning	Comments/experiments
pGEX4T2 Vps29	Gene was amplified from synthetic template (GenScript) and ligated using BamHI/XhoI restriction digestion sites	protein crystallization
pGEX4T2psp Vps29	Introduced additional cleavage site for PreScission protease	co-expression and purification of complex
pET28aH_6 Vps35	Gene was amplified from synthetic template (GenScript) and ligated using NcoI/XhoI restriction digestion sites, His-tag introduced with amplification primers	co-expression and purification of complex
pET28aH_6 Vps26	Gene was amplified from synthetic template (GenScript) and ligated using NcoI/XhoI, His-tag introduced with amplification primers	co-expression and purification of complex
pGEX4T2 Vps5	Gene was amplified from synthetic template (GenScript) and ligated using BamHI/XhoI	liposome pelleting
pRSFDuetH_6 Vps5	Gene was amplified from synthetic template (GenScript), cloned into MCS1, NcoI	dimerization, complex purification.

© 2018 Springer Nature Limited. All rights reserved.

TECHNOLOGY FEATURE

CRYO-ELECTRON MICROSCOPY SHAPES UP

As the imaging technique produces ever-sharper protein structures, researchers are racing to develop tools to assess how accurate they are.



LABORATORY OF NEUROPHYSIOLOGY AND BEHAVIOR/THE ROCKEFELLER UNIVERSITY

Structures of ion channels, such as this insect smell receptor, are tough targets for X-ray crystallography. Cryo-electron microscopy has revealed this and more.

BY MONYA BAKER

Once derided as 'blobology' for its blurry images, cryo-electron microscopy (cryo-EM) is now churning out high-resolution structures of everything from virus particles to enzymes. The number of cryo-EM images uploaded to the Electron Microscopy Data Bank (EMDB) has boomed from just 8 in 2002 to 1,106 last year — the same year the technique won its developers the Nobel Prize in Chemistry.

The quality of cryo-EM images now rivals that of X-ray crystallography, long the dominant technique for solving protein structures. The technique has also succeeded where crystallography has struggled: showing, for instance, how temperature-sensitive ion channels work, characterizing pathological proteins in neurodegenerative disease and detailing how viruses can interact with antibodies[1]. Consequently, many veteran crystallographers are giving up on crystals and freezing proteins for cryo-EM instead.

Publications of cryo-EM structures are coming in fast (see 'Widening the bottleneck'). But, some researchers worry, not everyone knows how to evaluate them, and some are calling for new practices and tools to help them do so.

CRITERIA CREATION

The surge in cryo-EM is largely a result of better electron detectors and image-processing techniques, says Richard Henderson of the Medical Research Council Laboratory of Molecular Biology in Cambridge, UK, who shared Nobel prize. But, he says, the field still lacks the kind of standardized tools for producing robust structural models that crystallographers developed as their field matured. "This has led to a lot of sloppiness," he says. "What is needed now are better criteria to encourage researchers to put more work into their model-building."

Instead, the race is on to publish structures with ever-better resolution, and that is discouraging careful work, says Holger Stark, an electron microscopist at the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany. Some published structures depict atom-level precision without acknowledging that certain regions of the structure are "fantasy", with scant data to back up any particular interpretation, he says. "It's just noise in areas where people have put in atomic coordinates."

There is no question that cryo-EM has enabled fantastic discoveries and that many structures are solid, says Gabriel Lander, a structural biologist at the Scripps Research Institute in La Jolla, California. But he cautions that many researchers are too quick to assume that all the details in the structure are correct. As a result, someone who uses a structure to design mutant versions of a protein to understand its mechanism, or who sees a ligand binding in a poorly defined spot, could end up doing ▶

© 2018 Springer Nature Limited. All rights reserved.

months of failed experiments, he explains. "I don't want the reputation of cryo-EM sullied by over-interpretation."

RESOLUTION FIXATION

Protein structures are often judged by a single factor: resolution, the level of detail a structure shows. That metric is straightforward to ascertain in crystallography, but not in cryo-EM.

In crystallography, a highly ordered lattice of tightly packed molecules is rotated through an X-ray beam, and the resolution of the resulting image can be calculated directly from the diffraction patterns made by the deflected photons. Those patterns are then transformed into 'maps' of electron density, which researchers combine with the protein sequence to build a model. The model represents how specific chemical building blocks of a protein fold into sheets and helixes (visible at a resolution of around 5 ångstroms), and how side chains of amino acids are positioned (which start to become visible around 3.5 Å). Big, floppy objects tend not to form ordered crystals, so as a rule, the smaller and more rigid the protein, the more amenable it is to crystallography.

In cryo-EM, proteins and other macromolecular complexes are flash-frozen in a thin layer of water, ideally not much thicker than the protein itself. Irradiating that layer with low-energy electrons produces 2D images of individual particles on the detector — fuzzy shadows cast from scattered electrons (see 'Modelling in ice'). Thousands or even hundreds of thousands of these noisy images are then computationally sorted and reconstructed to create a 3D map. Finally, other types of software fit the protein sequence into the map to create a model. The smaller the object, the noisier the images, so cryo-EM tends to work best for larger structures.

To avoid mistaking noise for signal, researchers typically split particles into two subsets and build 'half maps' from each. The correlation between those two maps is used to calculate resolution — but it's an imperfect proxy, says Edward Egelman, a structural biologist at the University of Virginia in Charlottesville. "That's not measuring resolution, per se, it's measuring consistency." And the resulting values, he says, must be taken with a grain of salt. Indeed, he says that the race to claim high resolution has sometimes led researchers to "silliness" — such as reporting resolution to a hundredth or even a thousandth of an ångstrom, a level of precision that makes no sense with cryo-EM.

Also, not all false signals are random noise. Egelman has demonstrated that systematic artefacts (such as computationally adding non-existent cylinders into both half maps) can drastically (and erroneously) improve the apparent resolution of a structure[2].

Sometimes researchers actually back-compute an electron-density map from the structural model that created it, and then revisit their data to select particles that are most likely to confirm the model. "It's a kind of bias," says crystallographer Piotr Neumann at the University of Göttingen in Germany. "This kind of cheating is not acceptable, but it's okayish." Another, more common, technique is to create a 'mask' of the expected overall shape of the protein and use that to exclude portions of images. Done judiciously, this boosts the signal-to-noise ratio; done aggressively, it shoehorns or 'overfits' data.

TWEAKED TO FIT

Structural biologists joke that there are many more structures published with resolutions of 2.9 Å than of 3.0 Å — an apparent symptom of over-aggressive analyses. But even without gaming, describing a protein with a single number is problematic, says Lander. It obscures the fact that the quality of a cryo-EM map varies dramatically, with the poorest-quality fit often occurring in the most flexible and biologically interesting areas of the protein. "There is no one metric that is good," says Neumann. "All metrics can be biased or not fully reliable. So, we need to use many simultaneously."

Earlier this year, Neumann and his colleagues set out to document how well protein structure models in the Protein Data Bank fit the corresponding maps in the EMDB. They found only low or moderate agreement for more than three-quarters of the 565 structures examined, suggesting that large swathes of the models should be viewed with scepticism[3].

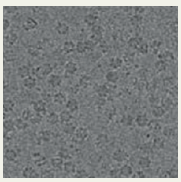
Some drug developers, at least, are approaching the models with caution. Christian Wiesmann, head of the cryo-EM team at the Novartis Institutes for Biomedical Research in Basel, Switzerland, says that when looking at models of proteins bound to small molecules, he typically downloads maps from the EMDB, assesses how other researchers nestled the compounds into the protein and then uses his own judgement. More than once, Wiesmann says, he would have made different calls — differences that could affect drug design.

Not every researcher possesses that level of structural sophistication. But even if they did, maps can be hard to vet. Authors must deposit their maps in the EMDB when publishing papers, but these deposits are often insufficiently annotated, says Alex Wlodawar, a structural biologist at the US National Cancer Institute in Frederick, Maryland, who has compared crystal and cryo-EM structures at high resolutions and found that the latter are often "optimistic"[4]. Researchers might deposit the raw map without the refined or 'sharpened'

MODELLING IN ICE

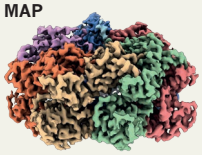
In cryo-EM, thousands of raw electron-microscopy images are collected and computationally analysed to build up a density map that reflects the shape of the protein.

RAW IMAGE



Where to share data
Electron Microscopy Public Image Archive (EMPIAR)

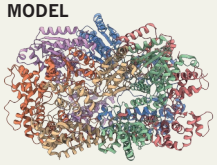
MAP



Electron Microscopy Data Bank (EMDB)

This is then combined with the known protein sequence to create a final model showing the placement of atomic groups.

MODEL



Protein Data Bank (PDB)

SOURCE: C. PUCHADES *ET AL. SCIENCE* **358**, EAAO0464 (2017)

Widening the bottleneck

Better validation tools and practices are not the only thing holding back cryo-electron microscopy; many researchers are unable to produce images of sufficient quality to even start the process.

"The real bottleneck is specimen prep," says Bridget Carragher, electron microscopy co-director at the New York Structural Biology Center. But a technique she has co-developed might substantially broaden that bottleneck.

Carragher and her co-director Clint Potter developed robots that have droplet-sensing cameras and piezoelectric devices similar to those found in an inkjet printer. The robots can apply small volumes of sample onto a grid just as another robot plunges the sample swiftly and evenly into cryogen, creating a thin, uniform, frozen layer that is perfect for imaging. The grid itself is covered in nanowires that wick away excess solution in a more controlled way than the blotting paper typically used. That also reduces the chance of proteins getting stuck at the air–water interface, which can cause them to denature or adopt preferred orientations rather than the random orientations that are necessary for them to be seen from every angle.

The system, called Spotiton, allows users to prepare more samples more quickly and using less protein, and ensures that a greater fraction will be usable. The pair have licensed Spotiton to TTP Labtech in Melbourn, UK, which plans to commercialize the system in the next year or so under the name Chameleon. M.B.

© 2018 Springer Nature Limited. All rights reserved.

map used to build the model, or without reporting whether they used a mask in building it. And very few deposit the half maps used to validate their analysis.

MAPPING THE FUTURE

Like models, maps are highly variable in quality, says Ardan Patwardhan, who manages the EMDB. Suites of automated and semi-automated tools have been created to help researchers turn 2D cryo-EM images into 3D maps. To help assess these workflows, the EMDB has run several validation competitions. It found that the greatest variability came not from the software packages, but from the experience level of the users. Less-experienced groups used default parameters; the best teams tailored settings to the data they had. That can make the difference between clearly visible side chains and blurry secondary structures, even when starting from the same raw images[5].

Today, researchers are calling for better methods for validating cryo-EM maps and models[6] — and raw image data could help. In 2014, Patwardhan and his colleagues at the European Bioinformatics Institute (EBI) in Cambridge, UK, created the Electron Microscopy Public Image Archive. The largest of the current 175 deposits of raw image data is more than 12 terabytes, which takes about 5 days to download.

Better methods for representing uncertainty could also help. Lander has proposed that researchers provide a spectrum of models[7] to better illustrate the range of structures that might fit the data. Maya Topf, a computational structural biologist at Birkbeck, University of London, has helped to create software called TEMPy that measures the quality of the model at the scale of amino acids rather than of the entire structure. Although this is not yet mandatory, the research community is starting to expect these kinds of evaluations, she says[8]. "The awareness is growing. More and more people are reporting in papers the local resolution."

Still, cryo-EM has a long way to go to match practices of crystallography. "The fact that data and models need to be validated has to become ingrained in people's minds, especially as the field attracts many new practitioners who don't have decades of experience," says Gerard Kleywegt, a structural biologist at the EBI. And, of course, some things are fundamentally different: crystallography captures proteins in rigid conformations, whereas cryo-EM can show more natural, and naturally ambiguous, conformations for which people are still developing the language to describe. Improvements will require better methods, greater consensus and better practices — all of which take time to develop. A validation task force met in September 2010 to develop recommendations, Kleywegt notes. "The field has evolved so rapidly since then that a follow-up meeting is overdue." Planning for a 2019 meeting is already under way. ■

Monya Baker *is an editor at* Nature.

1. Renaud, J. P. *et al. Nature Rev. Drug Discov.* **17**, 471–492 (2018).
2. Subramaniam, S., Earl, L. A., Falconieri, V., Milne, J. L. S. & Egelman, E. H. *Curr. Opin. Struct. Biol.* **41**, 194–202 (2016).
3. Neumann, P., Dickmanns, A. & Ficner, R. *Structure* **26**, 785–795.e4 (2018).
4. Wlodawer, A., Li, M. & Dauter, Z. *Structure* **25**, 1589–1597.e1 (2017).
5. Heymann, J. B. *et al. J. Struct. Biol.* https://doi.org/10.1016/j.jsb.2018.08.010 (2018).
6. Afonine, P. V. *et al. Acta Cryst. D* **74**, 814–840 (2018).
7. Herzik, M. A. Jr, Fraser, J. & Lander, G. C. Preprint at bioRxiv https://doi.org/10.1101/128561 (2017).
8. Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. *J. Struct. Biol.* **199**, 12–26 (2017).

Big data fresh from the sea

Machine learning helps marine biologists to churn through millions of plankton images.

BY JEFFREY M. PERKEL

When they think about big data, most researchers probably imagine genomics, neuroscience or particle physics. Kelly Robinson's data challenge involves plankton.

"A lot of things that we enjoy seafood-wise — from fish to oysters to mussels to shrimp — almost everything starts their lives as plankton," says Robinson, who studies marine ecosystems at the University of Louisiana at Lafayette. In photographs, they look like floating specks of dust, and her research involves quantifying and mapping their distribution and predator–prey interactions. The problem is, she must do so in millions upon millions of images.

Robinson collects data by towing a remote-camera platform called ISIIS — the *In Situ* Ichthyoplankton Imaging System — behind a boat. ISIIS captures about 80 photos per second, or 288,000 images (660 gigabytes) per hour. For one project in the Straits of Florida, when Robinson was a postdoc, she generated 340 million pictures; a colleague working in the Gulf of Mexico generated billions.

"You start to learn about things that you never thought you would learn," Robinson says, "like the number of files that you can store on an individual computer. It's 30 million, by the way, on your regular PC." On her most recent cruise, Robinson sailed with 52 2-terabyte hard drives, which a student had to monitor and replace as they filled up. Someone then must get that collection to the university, convert the files to Linux formatting, and upload them to a server — a process that takes 24 hours per drive.



Kelly Robinson and her team scrutinize images.

The team uses machine-learning software to automatically pick out and identify objects in the images. But the algorithms must be taught what to look for — this is a starfish, that is a prawn. Such features are relatively rare in the water, so finding pictures for the training set takes time. Over two months, Robinson and her team manually sorted through 2 million images to find enough to feed the algorithm. "It's a little mind-numbing, but if you're under the gun you can do it," she says.

Naturally, the team is looking to optimize the process. Working with colleagues at Oregon State University in Corvallis, where she was a postdoc, Robinson is testing whether she could accelerate her work by processing the images on multiple video card graphical processing units (GPUs) running in parallel. She is also looking into cloud computing as an alternative to Earth-bound clusters.

But infrastructure goes only so far; what the team really needs, she says, is more people to crunch the numbers. Unfortunately, data scientists are in high demand, and industry jobs are lucrative. "We have a lot of turnover," she says. ■

Jeffrey M. Perkel *is technology editor at* Nature.

CORRECTION

The referencing in the Technology feature 'How to teach an old sequencer new tricks' (*Nature* **559**, 643–645; 2018) was incorrect. The correct version can be found online at at go.nature.com/2wmnhgc.

© 2018 Springer Nature Limited. All rights reserved.

GRIEF Confronting loss presents difficult choices **p.571**

SHARE Tell us your career story at **naturecareerseditor@nature.com**

SOCIAL Follow us on Twitter at **twitter.com/naturejobs**

DR CAMILLA NORD



Neuroscientist Rebecca Lawson forged an independent path through her postdoc positions.

CAREER PROGRESSION

Independence day

Postdocs must seek autonomy in an ever-changing career landscape.

BY CHRIS WOOLSTON

When Rebecca Lawson started her postdoc position in the lab of Geraint Rees, a neuroscientist at University College London, in 2014, she prioritized something very important to her and her future: Rebecca Lawson. "I was thinking about independence," she says. "It's an interesting time in your career. You're working for someone else, but you have to have your eye on the prize."

Before Lawson joined the lab, she had told Rees that she needed the freedom to pursue her own ideas, run her own projects and generally prepare for a future in which she would be in charge of her own lab. "From day one, I turned up with a list of things that I wanted to work on," Lawson says. She ended up tackling everything on that list and then some, always following her instincts and judgement to find the next step. Very rarely, she says, Rees would question her choices. "I stuck to my guns on a couple of occasions, but there wasn't any conflict," she says.

Postdocs have many reasons to seek some degree of independence. Future hiring committees want assurance that a postdoc can thrive beyond the shadow of their principal investigator (PI), and postdocs need to know how to publish papers and win grants on their own to survive. For Lawson and other postdocs, establishing an identity and track record of their own is a key challenge on the path to a science career. To achieve that goal, they have to navigate often-tricky power dynamics, consider the attitudes and intentions of potential supervisors before joining a lab, set boundaries and, in some cases, learn how to say no.

Any postdoc's quest for independence can be greatly complicated by the well-defined hierarchy in the lab, says Hugh Kearns, a lecturer at Flinders University in Adelaide, Australia, who specializes in the development and training of junior researchers. "The PI has all the power," he says. "The postdoc ▶

© 2018 Springer Nature Limited. All rights reserved.

is usually new and vulnerable. They hope that advisers will look after them, and most of the time that happens. But it's a master–serf arrangement," he says. Through the weight of a recommendation letter, he adds, the PI can also decide whether a postdoc has a chance to get a job afterwards.

For this and other reasons, Kearns says, it's crucial for postdocs to investigate the PI, the lab and the lab's culture before taking a position. By asking the right questions and checking the fates of past lab members, postdocs can learn whether they'll end up working solely in the service of a PI or whether they'll have the chance to pursue their own ideas and advance their own careers. He recommends asking about important issues such as ownership of data and authorship policies. But postdocs shouldn't settle for verbal reassurance alone. "You should do some surveillance as well," he says. "Check where postdocs end up on the authorship list, and where they go when they leave the lab. You're a researcher, so do your research. Once you've joined a lab, it's probably too late."

Lawson says she researched Rees's lab before meeting him, and tried to suss out his attitude towards postdocs and their futures. "Our first meeting lasted for hours and hours," she says. "He made it explicitly clear that he cared about my career development as much as he cared about the science that I was producing." Rees encouraged Lawson to speak to other members of the lab, an exercise that confirmed her gut feeling that the lab was a place where she could pursue her own ideas and prepare for her own future.

Lawson taught herself new techniques and supervised master's-degree students to independently advance her career development (see 'Teaching tips for postdocs'). She also secured a small grant from the Experimental Psychology Society in Lancaster, UK, even though she was already working in a fully funded lab. "I wanted to demonstrate that I could do it," she says.

Rees says that almost all of his postdocs hope to lead their own groups one day, and that he tries to give them opportunities to prepare for that future. "They need to be able to be independent thinkers, but also able to work in a team," he says. "It's a difficult balance to do your own thing but also learn to trust others and develop their skills."

Lawson's commitment to her own career paid off. In 2017, she won a Wellcome Trust Sir Henry Dale Fellowship worth more than £980,000 (US$1.3 million) over 5 years. And in 2018, she became an affiliate lecturer and the leader of her own lab at the University of Cambridge, UK, where she studies autism spectrum disorder and other developmental conditions. "I had all the skills in place to be a PI," she says.

Lawson's road to independence was smoothed by the fact that Rees had funding through a fellowship from the Wellcome Trust that allowed him to hire postdocs who weren't constrained by a grant that was tied to particular projects. She had faced a different situation when she accepted her first postdoc position, with Jon Roiser at University College London in 2011. In that case, she had to work on specific experiments that were spelled out in the grant. Still, she adds, "we had conversations about how I might pursue my own research interests. I've been lucky to have two bosses who had open discussions with me about my career."

In the United States, nearly half of all biomedical postdocs are working on project grants, according to a 2018 report from the National Academies of Sciences, Engineering, and Medicine in Washington DC (see go.nature.com/2mtrzcp). The report notes that PIs on project grants have wide discretion about how much freedom and support to give postdocs.

To help more postdocs to transition to independence, the academies recommend significantly increasing the number of available training grants and individual fellowships. In addition, they call on the US National Institutes of Health (NIH) to add a new section to project-grant applications that would require PIs to spell out their commitment to the "scientific and professional development" of all postdoctoral trainees.



CHEYNE MAYES

Isaiah Hankel runs Cheeky Scientist.

Independence has been a top priority for Lauren Robinson, an animal-welfare postdoc working on a project grant at the New College of Florida in Sarasota; she is currently also a visiting academic at the University of Edinburgh, UK. She says that she pursued multiple affiliations because she wanted opportunities for research and collaboration outside her specific postdoc duties, which mostly involve tracking behaviours and hormones of zoo animals. To that end, she's also editing a book on primate welfare for Springer, a division of Springer Nature, which publishes *Nature* (*Nature*'s Careers team is editorially independent of its publisher). "I won a photo contest with Springer, and worked it into a book," she says. "My postdoc research doesn't focus on primates, so that project has given me a lot of independence."

Taking on tasks that go beyond the basic requirements of a position can help a postdoc to build an identity and reputation outside the lab, but finding the time to do so can be a struggle. "I feel pressure to say 'yes' to everything," Robinson says. "But I have to think about taking on too much. I believe a healthy balance produces better science."

For their own sake, postdocs have to learn how to say no, a word that marks a big step towards independence. But that can be tough, especially if the request comes from a PI, says Isaiah Hankel, a cell biologist who is founder and chief executive of Cheeky Scientist, a PhD-only career-advice business in Liberty Lake, Washington. "It's so hard for a postdoc to say no to a PI," he says.

Although it is challenging, Hankel recommends setting boundaries early in the relationship, telling PIs from the outset that the postdoc will need time for outside projects and career-development events. Postdocs should also try to get some control over their schedule from the

A MENTOR'S TOOLBOX

Teaching tips for postdocs

Mentoring junior scientists helps postdocs to demonstrate their independence and build transferable skills. Hugh Kearns, a science-career specialist at Flinders University in Adelaide, Australia, urges postdocs to work out mentoring expectations and responsibilities carefully with their principal investigator (PI) before taking on students. Here is some of his advice:

- If you're going to work closely with a student, offer to be involved with the recruitment.
- Clarify how much responsibility you'll give the student and how much credit you'll get for their work.
- Discuss how much supervision you're expected to provide.
- Discuss the types of problems that might arise — including interpersonal conflicts and poor performance — and how you might address them.
- Clarify your authority to make important decisions about research designs and budgets if the PI is unavailable for an extended period.
- Ask if there are options for receiving supervision training. **C.W.**

© 2018 Springer Nature Limited. All rights reserved.

beginning, he adds. "If you go into a lab and start working 16 hours a day, anything less than that will seem like slacking off," he says.

Kearns agrees that postdocs take a risk when they turn down a PI's request. "It's almost impossible for them to say no without feeling like they're going to get into trouble," he says. He encourages postdocs to harness the power of a simple phrase: "other commitments". As in, "I'd love to do that for you, but I have other commitments."

A track record of independent thinking and actions can be extremely valuable for postdocs looking to move on to the next step of their careers, but that record might become distorted through conscious and unconscious bias. Lawson says that women can find it harder to be recognized for their achievements. "You hear conversations in corridors suggesting that their success was really to do with their boss or their supervisor," she says. And they can face a double standard. "If you work very collaboratively and do well, you can be accused of not being independent enough. But if you put your head down and focus, it can be seen that you're uncollaborative."

Lawson says she never felt her gender caused anyone to question her independence or accomplishments, largely because she had gone to such lengths to forge her own identity in the lab.

Some scientists speculate that biased views could be an important but underappreciated obstacle to the progression of women and people from under-represented groups in science. In the United Kingdom, for example, women make up about half the of PhD students in physical or biological sciences, but their representation drops at the higher levels of a career, in a phenomenon known widely as the leaky pipeline.

Many factors play a part in this much-discussed problem — but one could be that women have a harder time convincing granting bodies, review boards and potential employers that they deserve personal credit for their accomplishments. "Maybe it's more difficult for women, all other things being equal, to persuade appointment committees that they are independent," says Ed Bullmore, a neuroscientist at the University of Cambridge, who says that he has seen women get unequal treatment during the hiring process. "It may be one of those ways we are biasing the odds against women."

Bullmore thinks that scientists — including those on hiring committees — should accept and embrace the fact that nobody succeeds solely on their own.

"I don't feel that anything significant that I've done has been truly independent," he says. "It's essential for science for people to work freely with one another." ■

Chris Woolston *is a freelance writer in Billings, Montana.*

COLUMN

The price of grief

Confronting loss can present an impossible choice, says **Summer Praetorius**.

The night before an interview for my dream job as a palaeoceanographer, I talked to my brother for the last time.

The position was one I'd sought throughout my entire career — it aligned perfectly with my research interests and was my best shot at a permanent job near my husband's work.

But the call with my brother was like a waking nightmare. I thought I was listening to him die over the phone as he gasped for breath.

This wasn't my first confrontation with major loss. When I was five, my mother narrowly survived a terminal-cancer diagnosis. When I was in high school, my brother developed schizophrenia and alcoholism. During my undergraduate programme, my father died unexpectedly. And, while I was in graduate school, my mother had a stroke — a magnetic resonance imaging scan showed an aggressive glioblastoma that claimed her life within a few months.

I managed each of these events as best I could without letting them derail my career in science, although I longed to quit so I could shrink the widening fracture between the demands of academia and my heart. But the question would always arise: what then? So, in the end, I tethered myself to the solid rock of science and clung on tightly.

The morning after the call with my brother, the job interview was like a continuation of a bad dream. So incoherent were my thoughts that it was as if I were watching my body from above. I felt like my dreams for a career in science were evaporating. I had skirted the black hole of grief my whole life, only to collapse at this important moment.

After the interview, I returned to the lab where I was a postdoc and told a supervisor what had happened. He suggested I contact them and say I hadn't been performing at my best, but he cautioned me to offer no excuses.

Practical advice, yes. But I think this is one reason junior researchers leave science — the demand to partition ourselves into separate entities can fragment our psyches.

Grief is like a hurricane sweeping through our brains; it can carve a fresh scarp through our self-confidence, leaving a fog in its wake. So many young scientists are already hanging on by a tenuous thread in an environment where there is no room for faltering.

This is especially true for those most susceptible to impostor syndrome — women, people of colour, anyone who belongs to an under-represented group. For those who have fought



ADAPTED FROM DANE MARK/GETTY

against voices telling them that they don't belong, this 'brain fog' might be perceived as proof that their dislocation in science is substantiated. Grief can be the tipping point that pushes young researchers into a false concession of their inability to hack it in science. This is echoed in a 2017 study that identifies a disproportionately high rate of mental illness among PhD students — especially those dealing with work–family conflicts (K. Levecque *et al. Res. Pol.* **46**, 868–879; 2017).

The paradox is that many scientists are driven to improve the world for humanity, but the culture of science can be dehumanizing. We need to promote a culture that recognizes our humanity, where normal, human failure and struggle are not equated with academic ineptitude.

If you are a young scientist struggling with grief, you might need time to sit with it. You might be forced to make difficult decisions. Be clear with yourself about what you are unwilling to give up, and forgive yourself the rest. You might find your goals changing on the other side of loss, but wait to make career-altering choices in the calm after the storm, not in the heat of heartache.

My brother died a week after our call. Two days after that, I was offered the job — a bittersweet victory. Science had ferried me to more stable ground, but demanded its fare in return. I am haunted by all that I couldn't give my family members in their final days. But this can be the bitter choice when confronting loss as an early-career scientist — escort a loved one to their death or keep your own dreams afloat. I wish it did not have to be so stark. ■

Summer Praetorius *is a research geologist at the United States Geological Survey in Menlo Park, California.*

© 2018 Springer Nature Limited. All rights reserved.

LAST CONTACT

A missed appointment.

BY GRAHAM ROBERT SCOTT

Lagos sits upon the sea.

Or, more precisely, it perches on the stacked, damp ruins of itself. Buried under layers of buildings lie corpses of former islands and reefs. A hundred years ago, they poked above the waves; now they don't.

Through the streets and Fuji-shaped arcologies of the city's upper layers flickers the Lagos Consciousness. The Consciousness calls herself *Nanshe*, after the Sumerian goddess of civilization and fishing, but it's been several years since anyone flesh and blood addressed her by that name. As a municipal consciousness, Nanshe manages resources affecting the agglomeration's staggering population — depending on who's asked and where they draw boundaries, between 76 million and 90 million people.

Nanshe's an AI-2 — a hive mind of smaller, specialized AI-1s, each designed to take over when a problem arises in its wheelhouse. Like, for instance, when a Category-2-protected western lowland gorilla appears out of nowhere to sit on a congested downtown roundabout.

Like, for instance, right now.

Deep within Nanshe, a wildlife-protection module stirs. Her traffic AI-1 cedes control, and for the first time, Nanshe's wildlife component runs the entire city.

She realizes almost immediately she's on her own. Contact permits haven't been issued for apes in decades. The last remaining holder, a woman who oversaw translocations to Nigerian reserves, died in Lagos five weeks earlier. Which means no human may touch the gorilla. Nanshe reroutes traffic and messages everyone in the district a legal warning not to interfere. Then thinks.

Proposition: no gorilla simply wanders into the middle of a city. Food, water, mates all lie in another direction. The city smells and sounds like danger. Either the gorilla escaped from local (illegal) captivity, or it's braved this environment for something specific. Nanshe orders her police-dispatch AI-1 to work on the first hypothesis, but the only way to test the second is to learn more about the gorilla.

An experiment: can Nanshe use facial recognition on an ape?

Indeed, she can.

The gorilla has a name: Leo leo panthera. The odd name dates him. Reserve experts only name animals after extinct species and clades, like the West African lion, when the



ILLUSTRATION BY JACEY

animal's own species is critically endangered. It's a rhetorical move. Watching videos about an endangered elephant, a girl learns he's named Dodo. Now, whenever she talks about Dodo, she reminds every adult in earshot about the species already in the ground. Could they donate a few bucks to save Dodo the Elephant? Why, sure, they could.

Although still protected, the western lowland gorilla hasn't been critical since 2068. Conclusion: Leo was named more than 30 years ago, making him an old gorilla.

(On his roundabout, Leo grunts and loops in figures of eight, sniffing the air. Commuters take pictures from inside self-driving cars. Nanshe overrides their locks in case they stop being amused.)

The police module returns with a series of photographs and sensor readings, documenting Leo's infiltration of Lagos through an unfinished hypertube line to Kinshasa. Leo entered the city deliberately. He has an agenda.

Time for analysis. Nanshe tags everything she has about Leo, using likely synonyms for each term, dumping it all into one digital bucket. She pulls everything she finds about the sources of Leo's data, tags it and dumps it into digital bucket two. Then she clustermaps connections between the buckets.

At which point, the answer becomes immediate and obvious.

Gillian Lianne Yu. The last holder of a contact permit. The one who died five weeks ago. While she lived, Gillian had visited Leo every two weeks. For decades. Videos show Leo and Gillian playing, pressing foreheads together, stroking each other's hair, speaking in signs. In the final footage, Gillian has an oxygen tube up one nostril, medical patches up and down both arms. She cries as they part. *I love you*, she signs. Huffy and confused, Leo signs back: *We play longer next time.*

Now she's dead, and Leo is in the city, peering into commuter vehicles.

Gillian has missed two playdates.

Leo freezes on the roundabout when he hears Gillian's voice.

"How's my lion doing this morning?"

He spins, seeking.

Nanshe switches to a speaker half a block away.

"Looking healthy, kid. Wanna go for a walk?"

Leo vaults the hood of a car to follow.

"Oh, Leo," dead Gillian says, "how'd you get this scar? Have you been fighting the other boys?"

He whips his head around in frustration, sees one of the speakers. Gives it a close inspection, then a brutal side-eye.

"Over here, Leo," Gillian says, farther down the street.

He thumps the speaker next to him, snorts. Grudgingly follows the disembodied voice. He has no other clues. Vocal breadcrumbs lead him to a 20-storey cemetery, each garden level open to the air, like an old parking garage. From there, he follows a speaking traffic drone to the fourteenth floor.

Leo stops when he recognizes the out-of-breath woman who filmed his videos. Gillian's daughter Maduenu leans on a tombstone, face sheened with sweat.

Dead, she signs.

Leo grumps. *Gillian mad.*

No. Dead. Here. Maduenu points. Pats the soil.

Leo approaches, inspects the patch of ground. It bears no resemblance to his friend. He sniffs. Touches earth. Picks at grass on her grave, grooming like it's hair. Signs *Sad.*

He reaches out to Maduenu for a hug.

Maduenu shakes her head. She has no contact permit.

Leo huffs. Spreads his arms wider.

Maduenu hides her mouth with one hand.

"I can't do this," she says. Sniffing, she shakes her head again.

Leo turns his back on her, sending a message. When he hears Maduenu sobbing, he knows she's received it.

Then Leo lumbers away. Nanshe keeps the traffic back, and sees him home. ■

Graham Robert Scott *teaches writing at Texas Woman's University. His stories have appeared in* Nature *and* Barrelhouse Online.

© 2018 Springer Nature Limited. All rights reserved.